# Minimal Explanations untuk Prediksi Binarized Neural Network Menggunakan Abduction = Minimal Explanations for Binarized Neural Network Prediction Using Abduction

Kezia Sulami, author



---

Abstrak

Machine Learning (ML) sebagai bagian dari Artificial Intelligence (AI) telah membuat komputer mampu melakukan hal-hal yang membutuhkan kecerdasan manusia secara otomatis. Binarized Neural Network (BNN) merupakan arsitektur ML modern yang memiliki keunggulan yakni penggunaan memori yang efisien dan performa yang baik. Namun, seperti neural network pada umumnya, BNN juga merupakan black-box model yang memiliki kesulitan dalam menjelaskan prediksi yang dihasilkan. Penelitian ini menggunakan teknik abduction untuk memperoleh minimal explanations, dalam bentuk himpunan pasangan fitur dan nilainya, dari hasil prediksi BNN. BNN dimodelkan sebagai model Mixed-Integer Linear Programming (MILP) dan selanjutnya disederhanakan menjadi model Integer Linear Programming (ILP) yang merupakan bentuk formal agar dapat dilakukan teknik abduction. Hasil penelitian menunjukkan bahwa teknik abduction dapat digunakan untuk menjelaskan hasil prediksi BNN. Penelitian ini juga menerapkan teknik abduction untuk menghasilkan penjelasan subset-minimal pada hasil prediksi BNN untuk beberapa dataset.
.....Machine Learning (ML) as part of Artificial Intelligence (AI) has enabled computers to do things that require human intelligence automatically. Binarized Neural Network (BNN) is a modern ML architecture that has some advantages: efficient use of memory and good performance. However, like other neural networks in general, BNN is also a black-box model that has difficulties in explaining the resulting predictions. This research employs the abduction technique to obtain minimal explanations, that is a set of pairs of features and its values, from a BNN prediction. BNN is modeled as a Mixed-Integer Linear Programming (MILP) model and then further simplified into an Integer Linear Programming (ILP) model which is a suitable formalism for finding explanations using abduction. This research shows that the abduction technique can be used to explain BNN predictions. Furthermore, this research applies the abduction technique to produce subset-minimal explanations on BNN predictions for several datasets.