Query languages for neural networks

Martin Grohe [1], Christoph Standke [1], Juno Steegmans [2], and Jan Van den Bussche [2]

[1] RWTH Aachen University, Aachen, Germany
{grohe,standke}@informatik.rwth-aachen.de
[2] Data Science Institute, UHasselt, Diepenbeek, Belgium
{juno.steegmans,jan.vandenbussche}@uhasselt.be

We lay the foundations for a database-inspired approach to interpreting and understanding neural network models by querying them using declarative languages. Towards this end we study different query languages, based on first-order logic, that mainly differ in their access to the neural network model. First-order logic over the reals naturally yields a language which views the network as a black box; only the input–output function defined by the network can be queried. This is essentially the approach of constraint query languages, which means that evaluating queries would require the use of the notoriously complex algorithms for constraint query evaluation. For this reason, we mostly consider this approach to be a declarative benchmark. On the other hand, a white-box language can be obtained by viewing the network as a weighted graph, and extending first-order logic with summation over weight terms. This approach is essentially an abstraction of SQL, immediately giving us access to the many efficient algorithms for SQL query evaluation. In general, these two approaches are incomparable in expressive power. To resolve this we introduce *model agnostic queries*, which are queries that always have the same result for any two models that represent the same function. Under natural circumstances the white-box approach can subsume the black-box approach for boolean model agnostic queries; this is our main result. We prove this result concretely for linear constraint queries over real functions definable by feedforward neural networks with a fixed number of hidden layers and piecewise linear activation functions.