Abstract

This dissertation is concerned with techniques for inclusion of domain-knowledge into Deep Neural Networks (DNNs). We are primarily concerned with real-world scientific problems with the following characteristics: (a) Data are naturally graph-structured (relational), (b) The amount of data available is typically small, and (c) There is significant domain-knowledge, usually expressed in some logical form (rules, taxonomies, constraints and the like). Broadly, there are 3 different ways in which the domain-knowledge can be incorporated into a DNN: by changing the input representation, by changing the loss function, or by changing the model (structure and parameters). We propose techniques for the inclusion of domain-knowledge into DNNs that change the input representation. In particular, our principal contributions are as follows: (1) We study the inclusion of complex domain-knowledge into Multilayer Perceptrons (MLPs) using relational features and propositionalisation [LDG91]. We propose a utility-based stochastic sampling technique for drawing features from a large but countable space of relational features; (2) We propose a simplified technique called 'vertex-enrichment' for incorporating symbolic domain knowledge into deep neural networks that deal with graph-structured data, known as graph neural networks (GNNs); (3) We propose a systematic technique to incorporate symbolic domain-knowledge into GNNs using the method of inverse entailment [Mug95] available in Inductive Logic Programming (ILP); and (4) We construct a sequence generation system using a modular combination of two deep generative models and a discriminator model based on (3), and use this system for a problem of early-stage lead discovery in drug design. Our implementations are techniques that combine neural networks and symbolic representations, resulting in new neuro-symbolic models, such as: Deep Relational Machines (DRMs), Vertex-Enriched Graph Neural Networks (VEGNNs), Bottom-Graph Neural Networks (BotGNNs), and a modular end-to-end neuro-symbolic system for the generation of novel molecules for drug design. Our primary hypothesis is that inclusion of domain-knowledge can significantly improve the performance of a deep neural network. We conduct large-scale empirical testing of our hypothesis, using nearly 75 datasets in the broad area of drug discovery that consist of over $200,000$ relational data instances and with domain-knowledge containing about 100 relations. In all cases, our empirical evidence supports the primary hypothesis and encourages the inclusion of domain-knowledge into deep neural networks for prediction and explanation.