Towards Architecture-Agnostic Neural Transfer: a Knowledge-Enhanced Approach

Seán Quinn and **Alessandra Mileo**
Insight Centre for Data Analytics, Dublin City University
{sean.quinn, alessandra.mileo}@insight-centre.org

Abstract

The ability to enhance deep representations with prior knowledge is receiving a lot of attention from the AI community as a key enabler to improve the way modern Artificial Neural Networks (ANN) learn. In this paper we introduce our approach to this task, which comprises of a knowledge extraction algorithm, a knowledge injection algorithm and a common intermediate knowledge representation as an alternative to traditional neural transfer. As a result of this research, we envisage a knowledge-enhanced ANN, which will be able to learn, characterise and reuse knowledge extracted from the learning process, thus enabling more robust architecture-agnostic neural transfer, greater explainability and further integration of neural and symbolic approaches to learning.

1 Motivation and Background

Modern Artificial Neural Networks (ANN) can leverage large amounts of data to be trained to perform hard tasks such as recognising objects in an image or translating languages. The process they use is equivalent to a feature extraction with respect to the raw data and an optimisation goal. This process exposes the underlying compositional and hierarchical structure of the concepts contained within high dimensional data, but does not typically provide high level access to such structure or easily facilitate it's re-use in related tasks.

Unlike ANN's, humans learn by building a conceptual model of the world, which relies on the persistence of known concepts across tasks, and the ability to carry a reservoir of background knowledge across domains. The authors of [Lake *et al.*, 2017] argue that such a conceptual model is fundamentally incompatible with the purely connectionist learning of a neural network, as it may require the exploration of structural variations in the networks architecture, which goes beyond the capabilities of gradient-based learning in weight space. One of the key challenges in making more human-like artificial intelligence is incorporating these properties of structured learning into the neural network learning paradigm.

Our research aims to create knowledge extraction and injection mechanisms to allow the knowledge learned by a neural network during training to be accessed and transferred to another network in an architecture-agnostic way.

In addressing this challenge, we have been inspired by recent influential review articles within the Deep Learning community [Lake *et al.*, 2017; LeCun *et al.*, 2015] which call for new approaches to enhance deep representations with background knowledge. This is considered to be a key future enabler to significantly improving the ability of machines to learn new tasks faster and in a domain invariant way.

2 Related Work

Early work in [Maclin and Shavlik, 1996] introduced a method for injecting symbolic knowledge in the form of a propositional rule set into a neural representation and demonstrated that knowledge injection improved the networks performance on given decision tasks. More than 20 years later, authors such as [Tran and Garcez, 2018] share our core objective of combining background knowledge with neural learning and outline modern architectures towards accomplishing this. Contrasting these papers and the wider knowledge injection literature illustrates the vast differences which exist on how the task should be interpreted an accomplished, with most aiming to achieve success in a specific use case or scenario. No unified vision has emerged on how to achieve knowledge injection in the general case.

Statistical relational learning and probabilistic graphical models for knowledge injection constitute relevant approaches to this task. Statistical relational learning models can be trained on graph structured data and then used to predict new facts about the world, equivalent to predicting new edges in a graph. These methods have been shifting to increasingly neural based implementations in recent years, such as lifted relational neural networks [Sourek *et al.*, 2015].

Methods for knowledge extraction from neural networks are typically broken into three categories; Decompositional approaches, which seek to extract rules from the level of individual hidden and output units within a trained neural network. Pedagogical methods, which treat the network as a black box and model the relationship between inputs and outputs and finally, eclectic methods, which involve a combination of some aspects of both approaches. [Garcez *et al.*, 2001] and [Kumar, 2012] demonstrate two effective knowledge extraction experiments.

3 Scientific Approach

Our key hypothesis is that enriching neural representations with knowledge, which has been extracted from another trained network or provided by a human, will facilitate successful learning of new tasks with less training data. As illustrated in Figure 1 this research differs from traditional neural transfer in that (i) we seek to facilitate the injection of human generated knowledge as well as knowledge extracted from a trained network, and (ii) we do not seek to transfer layers of a trained neural network directly to another network but instead aim for architecture-agnostic transfer through the use of an intermediate knowledge representation, and (iii) both the knowledge extraction and the knowledge injection process are considered in the approach, thus they operate on the same knowledge representation formalism.

We do not aim to create a novel knowledge representation formalism in this research but rather to draw upon the body of literature, where a variety of knowledge formalisms have been used in knowledge injection and extraction scenarios [Besold *et al.*, 2017]. We plan to compare and assess the suitability of a number of such formalisms to identify desirable properties, including correctness, expressivity and complexity.

We initially focus on classification tasks, where we are given a source classification task A and a target classification task B where we wish to transfer knowledge from trained network A to untrained, or partially trained, network B. We will evaluate our methods by contrasting the performance and accuracy of our knowledge enhanced network B+ with a network B which has been trained on data alone. Disparity between B and B+ will be measured in two ways: (i) the difference in classification accuracy given the maximum amount of data and (ii) the classification accuracy achieved by each network given varying amounts of data.

4 Initial Investigation and Next Steps

Initial work focused solely on knowledge injection, with image classification as the use case. Here we used network dissection [Bau *et al.*, 2017] to identify semantically meaningful features in the deep layers of a partially trained Convolutional Neural Network. We then crafted some basic propositional rules based upon the features we identified in the networks layers with respect to the output neurons (the classes). These rules were then converted into a neural format (layers of neurons) and connected to the network using an approach highly similar to [Maclin and Shavlik, 1996]. This mechanism proved to be functional but impractical as it required the model to be significantly trained in order for semantically meaningful features to be present and did not have any practical applications due to the requirement of manually identifying features and hand crafting rules.

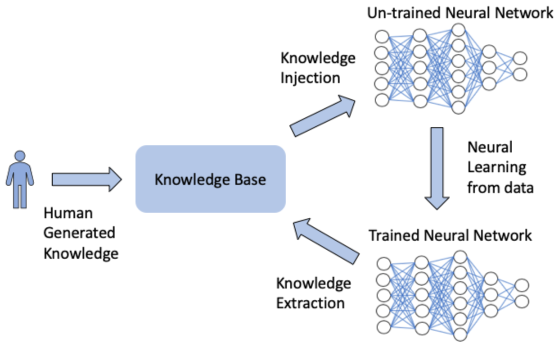
Un-trained Neural Network
Knowledge Injection
Neural Learning from data
Knowledge Base
Human Generated Knowledge
Trained Neural Network
Knowledge Extraction

Figure 1: High level overview of our proposed approach.

Current and future research aims to focus on graph-based knowledge representation formalisms and their associated models; probabilistic graphical models, graph convolutional networks and statistical relational learning techniques.

Acknowledgements

Funded by the Irish Research Council GOIPG/2018/2501 and partially by Science Foundation Ireland SFI/12/RC/2289. Supported by an Nvidia Corporation research hardware grant.

References

[Bau *et al.*, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

[Besold *et al.*, 2017] Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017.

[Garcez *et al.*, 2001] AS d'Avila Garcez, Krysia Broda, and Dov M Gabbay. Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, 125(1-2):155–207, 2001.

[Kumar, 2012] Koushal Kumar. Knowledge extraction from trained neural networks. *International Journal of Information and Network Security*, 1(4):282, 2012.

[Lake *et al.*, 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[Maclin and Shavlik, 1996] Richard Maclin and Jude W Shavlik. Creating advice-taking reinforcement learners. *Machine Learning*, 22(1-3):251–281, 1996.

[Sourek *et al.*, 2015] Gustav Sourek, Vojtech Aschenbrenner, Filip Zelezny, and Ondrej Kuzelka. Lifted relational neural networks. *arXiv preprint arXiv:1508.05128*, 2015.

[Tran and Garcez, 2018] Son N Tran and Artur S d'Avila Garcez. Deep logic networks: Inserting and extracting knowledge from deep belief networks. *IEEE transactions on neural networks and learning systems*, 29(2):246–258, 2018.