EE4080: Final Year Project Interim Report

Scientific machine learning for knowledge discovery

Academic Year 2021/22

Siddesh Sambasivam Suseela (U1822929B)

Supervised by Assoc. Prof. Mao Kezhi

Co-Supervised by Dr. Yang Feng

11 April, 2022

School of Electrical and Electronic Engineering

Contents

List of Figures

1 Introduction

1.1 Background

With an exponential growth in data and computing power, deep learning (DL) has been increasingly incorporated in numerous applications such as recommendation systems, natural language processing, drug discovery and image recognition, among few examples [2].

In specific, the growth has played a major role in increasing the research interest of using DL for modelling complex multidimensional systems. Conventional modelling of such systems is mostly based on using ordinary and/or partial differential equations (ODEs/PDEs) that govern a system's behavior [1]. These equations are usually derived from a first principal basis using axioms and/or existing proofs. However, there are **numerous unexplored systems whose governing equations are still unknown**. With advancement in data collection methodologies, rich representational data of such unexplored systems are being collected. As a result, **a data-driven approach has become a benchmark standard for knowledge discovery**, which aims to discover the underlying governing equations of a system directly from the observed data. Despite such refinements, knowledge discovery faces significant challenges such as noisy and corrupt data, latent variables, and the tendency for overfitting.

The aim of the project is to improve existing algorithms or propose a new machine learning algorithm to discover the governing equations of multidimensional systems. The following section briefly discusses the project scope and timeline.

1.2 Project Scope

The scope of the project are as follows:

1. Conduct a comprehensive literature survey and learn the state of the art (SOTA) algorithm in knowledge discovery and incorporation into AI/ML algorithms.
2. Implement and code various algorithms from existing work and publications.
3. Research on knowledge discovery and integration/incorporation in one of the following aspects,
   (a) Improve an existing algorithm

(b) Design a new algorithm

4. Incorporate the research work in applications through modeling equations for systems such as,

   (a) Burger's equation

   (b) Navier-Stokes equation

   (c) Nonlinear Schrödinger equation

1.3 Project Timeline

A project site has been used to centralize all the related work such as paper summaries, meeting minutes, research journal and other resources (Link to the site).
Phase 1: Conducting a comprehensive literature survey (Jan-May)
Phase 2: Identify specific aims of project based on your research vision, plan, preliminary data results and literature review results (Jun). **Phase 3:** Implementation and experimentation of proposed approach (TBC)

2 Literature Survey

The crux of scientific Knowledge discovery is to identify the underlying governing equation from a set of observational data. However, there are two main challenges to the task. First, instilling physics knowledge in neural networks so that the discovered equation obeys the known physical laws. Second, to introduce interpretability to the trained model, i.e to be able to parse the learned parameters to mathematical equations. Therefore relevant literatures to these major problems were studied for the project and the following subsections discusses the key ideas and limitations of a selective set of literatures.

2.1 Identifying governing equation of a system

2.1.1 Physics-informed learning of governing equations from scarce data

Key Ideas: The governing equation of a spatiotemporal system is generalised with a set of nonlinear, coupled parameterized PDE(s):

$$u_t + F[u, u^2, ..., \nabla_x u, \nabla_x^2 u, \nabla_x u \cdot u, ...; \lambda] = p$$

F is a nonlinear function; A deep neural networks essentially plays a role as a nonlinear functional to approximate the latent solution [1]. In order to mimic the above equation of the system, it is reformulated into the following form:

$$u_t = \phi\Lambda$$

Where ϕ is basically a function of u and Λ is a sparse matrix. The formulation is introduced as an additional term in the loss function.

$$L(\theta, \Lambda; D_u, D_c) = L_d(\theta; D_u) + \alpha L_p(\theta, \Lambda; D_c) + \beta \mid \Lambda \mid_0$$

$$L_p(\theta, \Lambda; D_c) = \frac{1}{N_c} \dot{U}(\theta) - \Phi(\theta)\Lambda \,{}_2^2$$

$$L_d(\theta; D_u) = \frac{1}{N_m} u^\theta - u^m \,{}_2^2$$

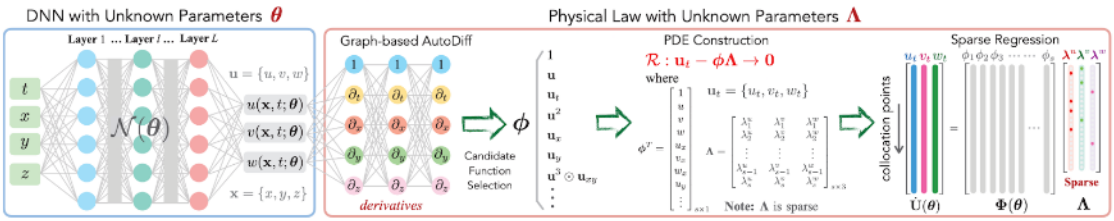

Figure 1: Fusion of PINNs, Sparse regression and optimization

The entire model is trained end-to-end with the introduced physics loss term and regularization of the sparse matrix (Λ). The overall approach is rooted in a comprehensive integration of bottom-up (data-driven) and top-down (physics-informed) processes for scientific discovery, with fusion of physics-informed deep learning, sparse regression and optimization. Results highlight that the approach is capable of accurately discovering the exact form of the governing equation(s), even in an information-poor space where the multi-dimensional measurements are scarce and noisy.

Limitations: However, a vast number of collocation point is required to maintain satisfactory accuracy and usage of automatic differentiation to model solution of higher dimensional systems results in computational bottleneck and optimization challenges.

2.1.2 Deep symbolic regression: Recovering mathematical expression from data via risk-seeking policy gradients

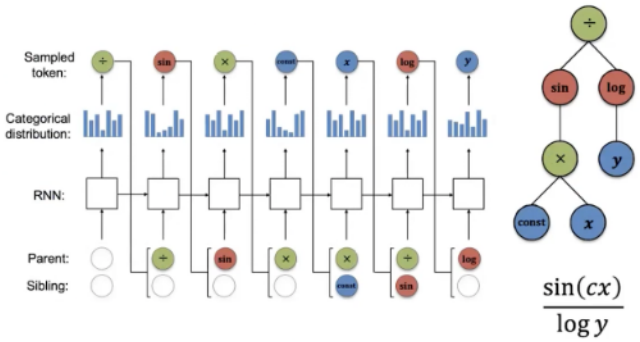

Figure 2: Sequence generation with RNN and symbolic tree

Key Ideas: The main construct of the paper is based on representing mathematical equations as symbolic trees and thereafter as a sequence of tokens (pre-order traversal) [5].

A Recurrent neural network (RNN) is used to search the sequence space for the mathematical equation which best fits the observational data. The training process is modelled like reinforcement learning (RL). The RNN acts as an agent and output of RNN as action. In addition, several rules are added to the outputs of the RNN which constrains the search space. The output is parsed to a symbolic tree and is optimized via risk-seeking policy. The reward function is the root-mean squared error (RMSE) calculated with the output and the observational data.

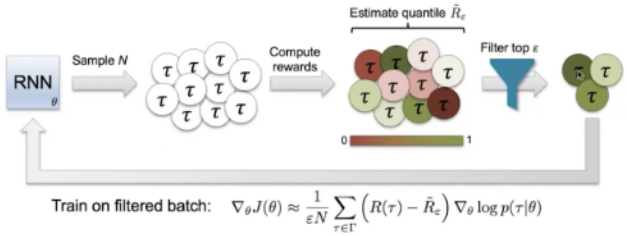

Figure 3: Equation filtering based on reward function

The deep symbolic regression method has shown great results in discovering mathematical equations. In addition, incorporating domain knowledge was eased into the training

processes as rules.

Limitations: However, the proposed approach overfits the the training data specially if size of the dataset is small and noisy. Furthermore, the model was not tested to predict governing equations of systems (e.g. Burger's equation, Navier-Stokes equation) but mathematical equations from observational data.

2.1.3 Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery

Key Ideas: The work by Kim et al. proposes a neural network-based architecture for symbolic regression called Equation Learner (EQL) network. The network is integrated it with other deep learning architectures such that the whole system can be trained end-to-end through back-propagation [3]. The building block of the approach is the symbolic layer (Figure 9).

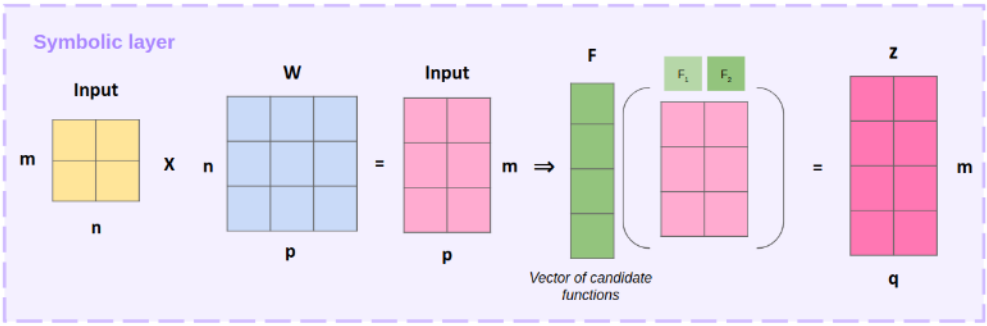

Figure 4: Overview of symbolic layer

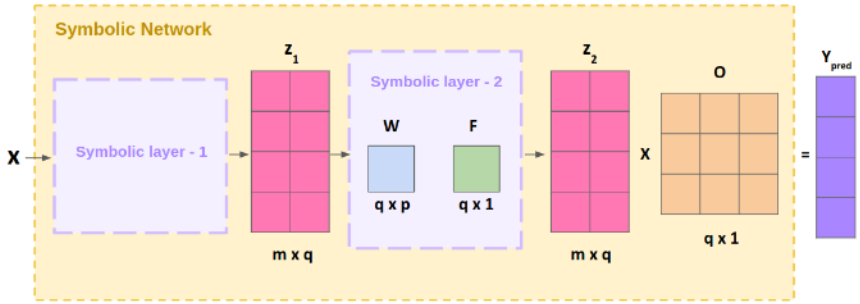

Figure 5: Overview of Equation learner network

Each symbolic layer consist of vector of functions instead of a single activation functions.

These functions are the candidate functions are parsed to form the governing equation of the system. Finally, the equation is parsed from the weights of the hidden layer using **sympy** python package.

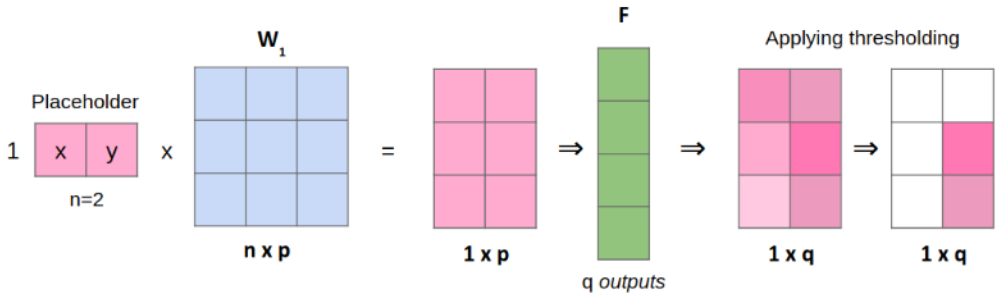

Figure 6: Parsing equations from hidden layer weights

Limitations: There are a few limitations to the proposed approach. First, the loss explodes a lot during the training process. Second, EQL network get stuck in the local minimum and fails to correctly identify mathematical expression. Finally, the network was not able to model high frequency terms with a limited dataset.

2.2 Instilling and modelling physics in multidimensional systems

2.2.1 Physics Informed Deep Learning: Data-driven solutions and discovery of Nonlinear Partial Differential Equations

Key Ideas: The paper proposes to train a neural network to solve supervised task subjected to law of physics described by a set of mathematical equations. An additional loss term is added to the loss function to penalize the network to account for the equations [6].

$$u_t + \nu[u; \lambda] = 0$$

ν is a non-linear operator parameterized by λ. This entire equation is added to the loss function and hence the model is optimized to minimize this equation.

Limitations: Although the approach is able to effectively instill the physics knowledge, it requires a high degree of domain expertise to formalize the loss function for a required task.

2.2.2 Physics-informed Convolutional Recurrent Network for Solving Spatiotemporal PDEs

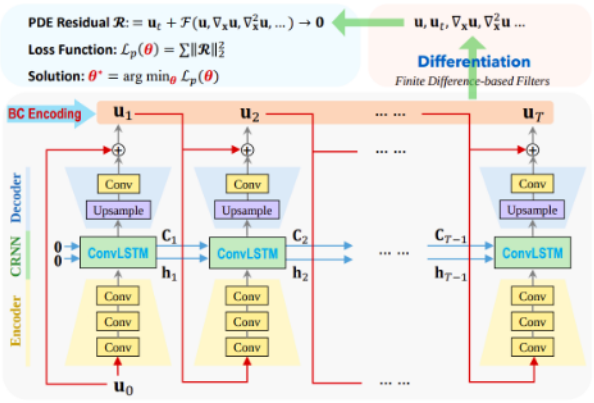

Figure 7: Model architecture of PhyCRNet

Key Ideas: Firstly, convolutional operators were used due to its faster convergence and better accuracy compared with fully-connected neural networks according to previous studies. Secondly, Recurrent units were used for controlling error propagation. It is important to note that the work focuses on regular (e.g., rectangular) physical domains, where both the spatial and temporal domains are discretized uniformly and convolutional filtering can be applied in nature [7].

Few key points to highlight are the following:

- convLSTM for modeling the long-period dependencies that evolve in time.
- Using Pixel shuffle layer to produce high resolution outputs from low-resolution feature maps.
- Introduces an encoder-decoder module, residual connect, auto regressive process and filtering-based differentiation.
- For improve the computational efficiency, the encoder is skipped periodically and computation is performed using a different input flow.
- a-RMSE is used as an evaluation metrics.
- PhyCRNet-s can generalize well given an initial condition for a specific partial differential equation.

Limitations: The current implementation can be extended to tackle irregular spatial domains by incorporating graph neural networks, as well as modified from forward Euler scheme to high-order difference strategy for more accurate temporal evolution modeling.

2.2.3 Physics-Informed deep learning for Scientific Computing

Key Ideas: The authors of the paper aim to implement an ensemble of neural network and classical/traditional linear solvers. The paper highlights a few important elements which can be extended to any experimentation processes [4].

1. The convergence behavior of the DL networks and PINN linear solvers is determined by the Frequency-principle (F-principle). Based on F-principle, deep neural networks (DNNs) often fit target functions from low to high frequencies during the training process.

2. In addition, it is found that having a large number of units at the beginning of the network and small number of units at the end of the network is detrimental to the PINN performance.

3. Initial hidden layers might be responsible for encoding the low-frequencies components and the following hidden layers are responsible for representing higher-frequency components.

4. The most impactful parameter for achieving a low training error is the activation function.

Limitations: The major limitation of the work is that only fully connected networks was used as surrogate network architectures. For solving the Poisson equation and elliptic problems in general, the usage of convolutional networks with large and dilated kernels is likely to provide better performance of fully-connected DL networks to learn non-local relationships a signature of elliptic problems.

2.3 Implementations Results

2.3.1 Implementation of deep symbolic network

The paper "Integration of Neural Network-Based Symbolic Regression in Deep Learning for Scientific Discovery" was implemented from scratch with PyTorch. GitHub Repository

Function	Predicted function	Number of layers	No. of epochs 1	No. of epochs 2
x	0.997938*x	2	10000	10000
x+y	0.995201*x + 0.995424*y	2	10000	10000
x+y+2z	1.00102*x + 0.996029*y + 2.00215*z	2	10000	10000
x*y	0.34948*(-x - 0.954117*y)**2 - 0.186232	2	10000	10000

Figure 8: Results of of the implementation

2.3.2 Implementation of Data-driven Physics Informed Deep Learning

The paper "Physics Informed Deep Learning: Data-driven solutions and discovery of Nonlinear Partial Differential Equations" was implemented from scratch with PyTorch. GitHub Repository

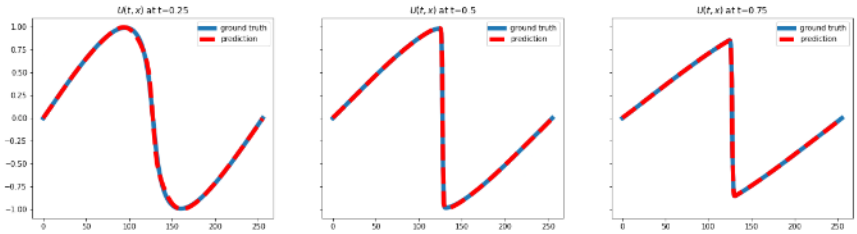

Figure 9: Modelling Burger's equation

3 Problems Encountered

Given the rapid advancements in the field, it is really challenging to keep track of the state of the art algorithms for knowledge discovery. In addition, for certain interesting papers, datasets were not provided hence was not able to replicate results for those work.

During the initial stages, I did not have access to GPU. So only after recess, I was able to successfully run experiments at scale. Finally, given the open-ended nature of the problem, there are so many literature that could be read regarding the topic, which made me feel overwhelmed. However, by restricting the number and type of papers to read helped to alleviate the problem.

4 Future Work

In the next two months, the plan is to understand the key challenges in application of deep learning in knowledge discovery. In addition, benchmark the discussed methodologies with datasets such as Burger's equation, Navier Stokes equation, among few examples. The current task is to create a comprehensive list of relevant research work exploring other methodologies in knowledge discovery.

5 Conclusion

Given the numerous applications of deep learning, its use in the field of scientific computing has profound implications. Specifically, being able to discover the governing equation of a system from a set of observational points could help us to understand several unexplored systems. On other hand, the ability of the model to express the learned parameters is a significant progress in interpretable AI.

The work done on the project includes studying existing literature to understand various methodologies and implementing state of the art algorithms in knowledge discovery. Hopefully, the research work done in project helps to improve the performance of existing algorithms or propose a new algorithm for knowledge discovery.

References

[1] Zhao Chen, Yang Liu, and Hao Sun. "Physics-informed learning of governing equations from scarce data". In: *Nature Communications* 12.1 (Oct. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-26434-1. URL: http://dx.doi.org/10.1038/s41467-021-26434-1.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[3] Samuel Kim et al. *Integration of neural network-based symbolic regression in deep learning for scientific discovery*. Aug. 2020. URL: https://arxiv.org/abs/1912.04825.

[4] Stefano Markidis. *The old and the new: Can physics-informed deep-learning replace traditional linear solvers?* July 2021. URL: https://arxiv.org/abs/2103.09655.

[5] Brenden K. Petersen et al. *Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients*. Apr. 2021. URL: https://arxiv.org/abs/1912.04871.

[6] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. *Physics informed Deep Learning (part I): Data-driven solutions of nonlinear partial differential equations*. Nov. 2017. URL: https://arxiv.org/abs/1711.10561.

[7] Pu Ren et al. *PhyCRNet: Physics-informed convolutional-recurrent network for solving spatiotemporal pdes*. June 2021. URL: https://arxiv.org/abs/2106.14103.