
Proceedings Paper:
Camilleri, D. and Prescott, T. (2017) Analysing the limitations of deep learning for developmental robotics. In: Biomimetic and Biohybrid Systems. 6th International Conference, Living Machines 2017, July 26–28, 2017, Stanford, CA, USA. Lecture Notes in Computer Science (10384). Springer, pp. 86-94. ISBN: 9783319635361. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-319-63537-8_8

Analysing the Limitations of Deep Learning for Developmental Robotics

Daniel Camilleri and Tony Prescott

Psychology Department, University of Sheffield,
Western Bank, Sheffield, United Kingdom
d.camilleri@sheffield.ac.uk
http://www.sheffield.ac.uk

Abstract. *Deep learning is a powerful approach to machine learning however its inherent disadvantages leave much to be desired in the pursuit of the perfect learning machine. This paper outlines the multiple disadvantages of deep learning and offers a view into the implications to solving these problems and how this would affect the state of the art not only in developmental learning but also in real world applications.*

Keywords: Deep Learning, Artificial Narrow Intelligence, Transfer Learning, Probabilistic Models

1 Introduction

Deep learning approaches currently dominate the scene of Artificial Intelligence (AI) and soon robotics. This approach to learning has been consistently yielding incredible results in multiple fields of AI. From object recognition [1] and object segmentation[2] to speech recognition[3] and even human-like speech synthesis with applications like WaveNet[4] and Deep Speech [5]. Yet, behind the illusion that more depth, more data and more computational power can solve any task, one can start to see the limitations that this approach to learning entails. As we shall see, depth carries with it a troublesome companion: obscurity in the understanding of *what* has been learned. Deep Learning has proven to be a sufficient and easy solution for limited, ad-hoc problems. However, the more complex a task becomes and the more flexibility an environment requires, the more obsolete feed-forward systems become in dealing with this challenge.

In this paper, we discuss the various disadvantages that deep learning methods exhibit. We start with the low transparency and interpretability of this approach and how these models require incredible amounts of data to train. Subsequently we see how the models created are very specialised and narrow and furthermore lack the ability to perform transfer learning between tasks. Thus the aim of this paper is to present a case for how different approaches are required to overcome these drawbacks in order to create machines that are capable of learning from experiences in a more neuromimetic manner.

2 Disadvantages of Deep Learning

2.1 Low Transparency and Interpretability

The information and experiences encountered by a robot can be stored in multiple forms of which two are the most commonplace: either in the form of Formal Symbolic Models or in the form of Sub-Symbolic Models. A Formal Symbolic model makes use of meaningful representations to express information in a compact manner. Instances of Formal Symbolic models include Hidden Markov Models and Bayesian Graphs. They are defined by graphs, where each node in the graph represents a whole. This makes the encoding of information in this way a tedious manual process which would need to define the features that are to be extracted from the input sensory streams. But it also makes the model explanatory in nature because given a series of activations within the graphs making up these models, one can trace these activations and thus discern how and why the model responded in the way it did.

On the other hand, Sub-Symbolic model approaches like Deep Learning represent information as parts that when combined in different ways make different wholes. Taking the example of a spoken word, a Symbolic Model would represent the whole waveform that describes the word while a Sub-Symbolic Model stores different features in different nodes such as the different power spectra making up the sound. This approach is powerful in that it can be trained automatically and will discover which patterns in the data are relevant to the required task and which are to be ignored leading to a really good understanding of the data at a very low level. However this understanding cannot be conveyed to the user or even to an expert because of the cryptic inner workings of a deep model. Thus once trained and optimised, sub-symbolic models are, for all intents and purposes, a black box. This results in two major disadvantages with the use of Sub-Symbolic models for robot learning.

The first is that the correct operation of the model cannot be formally verified. Dataset-centred verification is the only method of checking for correct operation under normal conditions and the operation of the model in abnormal conditions is covered in so far as abnormal conditions are covered within the training dataset. Therein lies the problem because the construction of an all-encompassing dataset is impractical and in reality unachievable.

To compound the issue, it has also been demonstrated that the current state of the art can be fooled into consistently making wrong decisions by what are called 'adversarial' inputs which is the second major drawback in this category. Adversarial inputs are inputs to a model which have been modified by a very small percentage in key areas, leading to a consistently wrong classification [6, 7]. In a real world application this is a serious vulnerability especially because the changes that create these adversarial inputs are small enough to be imper-

ceptible by humans.

This weakness combined with the lack of transparency presents an obstacle in the application of these models to certain real world applications which are legally required to demonstrate a high level of safety across all operating conditions like in the case of fully autonomous cars and drones [8] and in general any fully automated physical body including robots.

As an extension of this argument, the use of non-transparent models raises an issue whenever decisions taken by these models carry with them a level of required accountability. In the end, the users of these models are legally accountable for actions taken based on the prediction of these models and as such when queried about their decisions they are required to fully justify said decision. The absence of interpretability in the current state of the art renders their use, in financial situations like loan approval or healthcare applications like cancer diagnosis, problematic because they lack the explanatory power that humans have when reaching a decision.

There is also a human interaction component to the requirement of transparency and that is trust in the model's performance which ties in with accountability. The motivation for building more and more intelligent algorithms and models is to automate and improve and therefore we hold these models to a higher standard than we would a fellow human performing the same task. This, therefore, also demonstrates the requirement for transparency in trusting that a model is performing the correct decision.

Thus the ideal algorithm would exhibit the automatic pattern extraction properties and the classification accuracy of a sub-symbolic algorithm while still retaining the high level representations of a symbolic model. In this way, the model can be used to explain the 'reasoning' behind a decision in the manner that a symbolic algorithm is capable of.

2.2 Large Data Requirements

Deep learning, even in its application to narrow intelligence as we shall see later on, calls for the optimisation of millions of parameters via training. Consequently this leads to the requirement for large amounts of data in order to train a well-performing deep model. Furthermore, the data in most cases is required to be labelled which introduces an additional hindrance in the training of deep models although unsupervised approaches are also available.

These requirements mean that several problems which could benefit from the pattern classification power of deep learning, do not contain enough data to leverage this benefit. Data for particular applications like object segmentation or speech recognition are easy to collect and there are plenty of resources to obtain

this data. However, certain complex applications such as emotion recognition require information from multiple sources such as ECG, galvanic skin response and vision which is not easy to collect. Furthermore, due to the complexity of the inputs for the model, this further increases the amount of data required to train a good performing model. The same obstacle applies in industry with small businesses which do not generate enough data to train a meaningful model. Thus, these businesses cannot glean its benefits like larger businesses do. Consequently the collection of clean and abundant data is currently the biggest hurdle in the design of an AI system for both industry and academia.

Finally, since their learning is based on statistical processes, this type of learning cannot be used to remember one off events, like for example the event of purchasing your first car, something that humans do so well with the use of episodic memory. A plausible solution to this problem would be an architecture that becomes dynamically more complex as experiences are encountered by a robot and is then compressed when enough observations have been collected.

This would allow a robot to learn specific concepts quickly but also provide a means for performing generalisation when enough statistical data is collected. This generalisation would compress the complex experiential nodes into a simpler network that nonetheless preserves or even improves the classification accuracy of the uncompressed state.

2.3 Artificial Narrow Intelligence

Algorithms that learn have existed for more than 30 years in one form or another. Of interest is the fact that these algorithms have barely changed in that time, and the term learning is often used as a synonym of performance optimization. However, having robots present in highly variable environments like our homes, conference venues, or the streets, that can act robustly enough to be moved from an environment to another requires them to have the necessary mechanisms to acquire a vast array of diverse knowledge that includes grounding between concepts in different types of knowledge.

The current state of the art is however very good for Artificial Narrow Intelligence (ANI) applications: applications which have a very specific and narrow but deep knowledge of a certain task or process such as speech generation with WaveNet [4] or playing Go with AlphaGo [9]. Both of these applications are examples of models which have a very specific task (narrow area of knowledge) at which they have an incredible, sometimes super-human, performance because their knowledge of the area is very deep. However, what makes these models narrow, is that neither of these models can be used for anything other than the task that they are trained on. AlphaGo cannot be trained to play chess while still being capable of playing Go and WaveNet cannot be used to generate animal sounds while remaining capable of generating human sounds.

This means that ANI lacks the ability of generalising between different but similar applications and also lacks the ability of combining multiple separate sensory streams to perform multi-modal sensory integration. These properties are however quite important for the development of a learning robot because when combined they provide a manner of grounding one sensory experience within another and provided the basis for the creation of models that have less deep but much broader knowledge.

2.4 Transfer Learning

Transfer learning is the application of knowledge learned during the execution of a primary task to the execution of a new but similar secondary task. An example of transfer learning would be applying techniques learned from one game to a similar but different game which makes it a crucial skill for a robot to have in order to speed up learning across different tasks. However, due to the gradient-based nature of the algorithm, Deep Learning techniques are not by themselves capable of transfer learning. This has been demonstrated in [10, 11] where a network which is trained on a primary task and then trained on a similar secondary task can learn the second task much faster but at the expense of forgetting the first. This is called catastrophic forgetting.

Catastrophic forgetting is one of the primary research areas in the field of Deep Learning and a recent approach by Deep Mind [12] demonstrates the challenges inherent in achieving this and a possible solution which combines genetic algorithms with Deep Networks. This approach freezes the configuration of certain dominant pathways that are present within the network for the primary task in order to retain their functionality while allowing the rest of the network to learn normally. This is done in such a manner with the use of genetic algorithms such that it allows the learning of the second task as a function of the first, keeping the most fit pathways between different iterations of the genetic algorithm.

This has been the first demonstration of transfer learning within deep learning but due to its requirement of genetic algorithms is a trial and error driven approach to transfer learning requiring thousands of hours of training in order to achieve transfer learning between two tasks.

The difficulty of transfer learning derives mainly from the backpropagation algorithm which is central to all deep learning techniques and this is because the backpropagation algorithm essentially changes the representation of each node within the network in order to find the best features for the task at hand. This issue can be solved by having a fixed representation at each node but instead dynamically growing the network when new representations are required to handle a new problem.

Furthermore, the fixed representation of nodes within the network opens a new possibility for greatly accelerated learning when performed across multiple robots. Due to the fixed representations, a network learned in robot A can be matched and aligned with that learned by robot B and subsequently fused in order to combine the knowledge and experiences of the two into a single network. This can then be redistributed to A and B thus effectively making their learning rate significantly higher.

2.5 Probabilistic Models

Another issue with deep learning is that the models created are not fully probabilistic models. They can provide a probability vector of length n that describes how probable the current classification is with respect to the n classes that the model has been trained on. However this is not sufficient in a real world environment especially in the case of a developmental approach to learning. In these cases one must be able to also distinguish whether the current classification is something known or unknown to the model.

This capability is crucial for a learning machine because it defines the extent of knowledge for said machine. Given the capability of recognising known and unknown provides the machine with the capability for active learning where it can ask a human about the unknown object and thus expand its knowledge. Furthermore, a fully probabilistic model with the property of fixed node representations implies that the transitions of each node are assigned a probability that has been learnt from experience. This probabilistic representation can therefore be used to carry out predictions much like the associative action of System I thinking in the human brain [13]. This makes for a powerful system that conceivably brings us a step closer to implementing human intelligence in a machine.

2.6 Generative Models

Generative models are models that operate in two directions. They can go from a high dimensional input to a low dimensional classification or vice versa generate a typical high dimensional output given one of the classifications it has been trained on. The brain is a generative model due to the evidence we have for episodic replay where FMRI studies have shown the different parts of the brain involved in sensory processing light up when recollecting a memory [14]. Furthermore, using the human brain as inspiration, its generative connections allow it to perform what is called top-down control: where an activation in a high level network biases the functionality of a lower level network. A classic example of this is guided search where the features of an object that is being actively searched become more salient in the visual field. [15]

Deep learning methods have been applied to the field of generative models and the current state of the art are deep convolutional generative adversarial

networks [16]. These GAN operate by having two networks connected together in a cyclic form where the first network takes a high dimensional input such as an image and learns to classify that image. The input of the second network is then connected to the output of the first and its role is to take the classification vector received and generate from it a high dimensional output. This cycle can then be repeated several times in order to train the generative model. The issue with deep learning in this case lies in the complexity of the system which results in unstable training [17] when creating generative model.

3 Impact of Solving the Challenges

An approach that is capable of solving these challenges would create a great upheaval in the domain of AI and robotics much like the mainstream adoption of deep learning has but on a bigger scale due to the flexibility of a system without these disadvantages. The impact, would be incredibly far reaching and would mark the start of a new era of interactive learning machines. From fully autonomous transport, improved finance, advancements in the efficiency and accuracy of healthcare and last but not least the realisation of autonomous, embodied learning robots.

First and foremost, solving the issue of interpretability would mean that machine learning techniques can now be deployed in safety critical areas. This is not possible with current deep learning techniques but highly interpretable models provide the opportunity for formal validation and verification of operation under all conditions. Examples of safety critical areas include all forms of automated transport: autonomous cars, autonomous drones and even autonomous spacecraft where the risk is high but the option for failure is extremely low requiring reliable systems of guidance. The same applies for the use of robots in an industrial setting where robots work alongside humans.

Second, with the use of techniques that do not require extraordinary amounts of data one would see an increased adoption of these techniques. Furthermore when combined with interpretability this provides a method that is capable of analysing and extracting structure from small datasets that can be conveyed to the user. This is especially useful in the case of data analysis in small businesses in order to improve their product, an advantage which is currently reserved for only big companies with lots of data. This can even be extended to create a technique which is capable of storing information in a manner similar to the human brain with multiple hierarchies representing sensory, episodic and semantic memory. This provides an architecture that is capable of performing one shot learning on scarce unrepeatable data but at the same time capable of generalised learning.

Additionally, overcoming ANI and bypassing the issues with transfer learning in deep learning applications, one would create an algorithm that is capable of

applying knowledge learned in the solution of one problem to other problems via the process of analogy. Solving the problem of ANI would allow for the creation of robots capable of general developmental learning much like infants do in their early years by supporting multi-sensory integration and hence the grounding of concepts within different senses. But it is only with a solution for transfer learning that these robots can move past the early stages of human development and become intelligent machines.

Finally, with the use of fully probabilistic models that can recognise between activations that are known and unknown one can create machines that actively learn about their environment and also machines that are capable of interactive learning, where an unknown stimulus can be demonstrated to a human with the use of a generative property and then the human can provide a label to the stimulus thus driving the semi-supervised learning in the robot.

4 Conclusion

The approach of deep learning has provided a significant boost to the performance of models in different areas but contain the above problems which limit its incorporation in industry in safety critical applications and especially in robotics. At the same time, AI is becoming increasingly dominant and present in every aspect of our lives and its proliferation will continue to increase together with its power at solving increasingly difficult problems. Yet in order to have more interactive human-robot interactions, the robots need to be able to communicate in a human-understandable way and provide reasons for their actions. This is the focus of our future work which will look into a machine learning approach that performs as well as deep networks but at the same time does not have the outlined issues.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
2. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
3. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, IEEE (2013) 6645–6649
4. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. CoRR abs/1609.03499 (2016)
5. Arik, S.O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Raiman, J., Sengupta, S., et al.: Deep voice: Real-time neural text-to-speech. arXiv preprint arXiv:1702.07825 (2017)

Analysing the Limitations of Deep Learning 9

6. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS P). (March 2016) 372–387
7. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 427–436
8. Koopman, P., Wagner, M.: Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety **4**(2016-01-0128) (2016) 15–24
9. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587) (2016) 484–489
10. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of learning and motivation **24** (1989) 109–165
11. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211 (2013)
12. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. arXiv preprint arXiv:1701.08734 (2017)
13. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
14. Wheeler, M.E., Petersen, S.E., Buckner, R.L.: Memory's echo: Vivid remembering reactivates sensory-specific cortex. Proceedings of the National Academy of Sciences **97**(20) (2000) 11125–11129
15. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. Psychonomic bulletin & review **1**(2) (1994) 202–238
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)