Continual Learning: The Next Generation of Artificial Intelligence

DANIEL G. PHILPS

PREVIEW *Dan Philps provides an introduction to automated machine learning and its possible next-generation realization,* continual learning *(CL). CL advances the state of the art by attempting to automatically learn different tasks while retaining knowledge from previous model implementations. This article presents an application of CL to investment decisions. It also offers the interesting perspective that complexity is not simply a technical characteristic of a model formulation, but also a resultant of the application of human judgment. Although CL may be more technically complex than many forecasting models, it reduces if not eliminates the complexity from judgmental human inputs.*

INTRODUCTION

Go simple or go complex? For an applied-data scientist, simplicity wins every time (although with just enough "complexity-veneer" to hedge the next promotion).

There is considerable evidence that undue complexity reduces forecasting accuracy (Green and Armstrong, 2015). It also detracts from interpretability and costs more in time, tech, and resources. So how can it be that machine learning (ML), generally considered to be complex, presents forecasters with an unmissable opportunity?

Answer: while ML is generally perceived to be complex, it can actually reduce complexity in model development by avoiding human behavioural biases and by automating intermediate steps. In addition, if complexity serves to encompass a richer variety of information and if learning from this information can be automated, complexity is a price worth paying. For these reasons, ML has started to become a powerful resource for forecasters. This article discusses the potential benefits of two types of ML: automated (autoML) and continual learning (CL). Both can be described as end-to-end approaches, meaning they can directly convert input data into an output forecast, bypassing traditional intermediate steps.

ELIMINATING OUR OWN COMPLEXITIES

The subjectivity and behavioural biases we as forecasters tend to introduce to a modeling process are only partly tempered through experience. Biases include *confirmation bias* (Hergovich and colleagues, 2010) towards our latest favored approach; *cognitive dissonance* (Festinger, 1957), where we rework past errors to fit a competent perception of ourselves; and the *availability heuristic* (Tversky and Kahneman, 1973), where we bias our approach to cues that happen to be at the forefront of our minds. While the perception of ML is one of complexity, ML may actually be a way of automating away the greater complexity of our own inductive biases.

Following on from Spyros Makridakis's excellent series of articles in *Foresight* (Makridakis, 2017-18), this piece first describes automated machine learning (AutoML), and then what is likely to be the most disruptive end-to-end ML technology you have never heard of, continual learning (CL). This piece then explains why both are likely to become indispensable tools for forecasters.

AutoML

AutoML attempts to automate the steps an expert human would take to complete a forecasting project, thereby reducing

Key Points

- Automated machine learning could become an indispensable tool for forecasters, as it has for data scientists, but practitioners first need to take a pragmatic view of the perceived complexity/disadvantages of machine learning (ML).
- This article addresses how automated machine learning (AutoML) might mitigate forecasters' complexity concerns and introduces perhaps the next step in ML's evolution: *continual learning* (CL).
- CL has even greater potential to further automate forecasting by building a memory of different tasks over time, addressing what is called the catastrophic forgetting problem.
- The authors provide an example of how CL was successfully implemented in the real world to guide investment decisions.

the complexity in model development. AutoML's big advantage is that it allows forecasters to tap into the power of ML with minimum engagement in the underlying ML algorithms.

The original motivation for AutoML was to increase the productivity of researchers and reduce the probability of errors. However, AutoML has now exceeded these initial aims by becoming capable of learning from past operations (Feurer and colleagues, 2015). For instance, some commercial AutoML systems now learn to associate different shapes of input data—*metadata*—with preprocessing and model selection choices that have been effective in the past. This is described as learning to learn, or meta-learning.

Figure 1 displays an AutoML system; input data is passed in with a forecasting target specified (Ytest). The system then attempts to build an appropriate solution. First, *meta-data* is extracted (i.e. information that describes the input data) from which the system may attempt to guess options and settings to use. Many different learners are also tested, which can range from linear regression through to ARIMA, multi-layered perception, and support vector machines (SVMs). Gradient boosting trees and random forests are popular choices. The winning approaches (and associated settings) are generally chosen using fairly traditional statistical model selection combined with brute force grid searches. The most effective algorithms are shortlisted and then, typically, an ensemble of these approaches is formed to perform a final forecast.

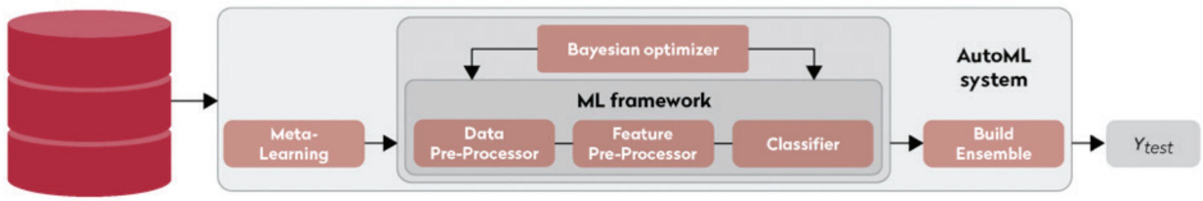
AutoML gives forecasters wide-ranging access to a broad toolbox of learners in a packaged pipeline—a neat way of consolidating existing algorithms and ML approaches while reducing the complexity of model development.

CONTINUAL LEARNING

While AutoML is a powerful tool, it is mainly just a consolidation of first-generation ML approaches. The next generation of ML, while drawing on similar building blocks, offers much greater potential. An example is continual learning—but is it mature enough to use in a forecasting process?

In ML, once a new model is learned, all previous models tend to be forgotten, an effect called *catastrophic forgetting*. In contrast, CL attempts to extract knowledge from a stream of information over time,

Figure 1. A Simple AutoML System



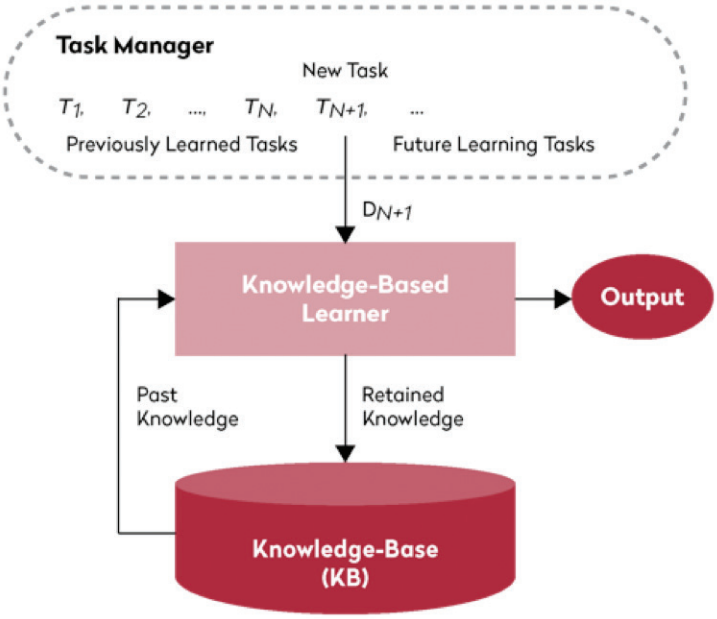
Source: An AutoML system for classification based on Auto-sklearn to then build memories (a *knowledge base*) that can be used to improve future forecasting, as illustrated in **Figure 2**. The idea is to have a system that generalises across different forecasting tasks (T_n) and perhaps modes of data, and retains the task-specific knowledge of *what works best* for each. This knowledge can then be recalled and applied if a similar task is encountered again in the future.

Research into CL commenced in the 1990s from a desire to construct knowledge-accumulating machines to avoid the catastrophic forgetting of traditional approaches (Silver, 2015). However, enabling machines to learn over time faces the serious challenge known as the *stability-plasticity dilemma* (Mermillod and colleagues, 2013), in that a system should learn over time but not at the expense of corrupting older knowledge.

A number of solutions have been proposed. In the late 1990s, Sepp Hochreiter and Jurgen Schmidhuber (1997) introduced the long short-term memory (LSTM) approach, which allows a recurrent neural network (NN) to forecast sequences—words in a passage of text, for example. (For a *Foresight* tutorial on neural network architecture basics, see Batchelor [2005], and for a more recent review of NN, see Januschowski and colleagues [2018]). Alex Graves and his team at DeepMind (2016) made a start in overcoming catastrophic forgetting, and subsequent researchers focused on how the extensive memories created by CL could be compressed into a knowledge base using *elastic weight consolidation* (Kirkpatrick and colleagues, 2017). However, there is only so much information you can squeeze into even a deep neural net before an information saturation point is reached.

An analogy of continual learning is how a child learns to ride a bike: wobbly at first, and then as skill develops with practice the neural pathways are reinforced and harden in the brain. Once learned, this skill is difficult to forget; in addition, it can be augmented if the child graduates to mountain biking, or *transferred* in learning to ride a unicycle. Simulating this effect with technology is challenging because the stability-plasticity problem must be addressed. Fortunately, real-world progress is being made through *continual learning augmentation* (CLA).

Figure 2. A Simple CL System



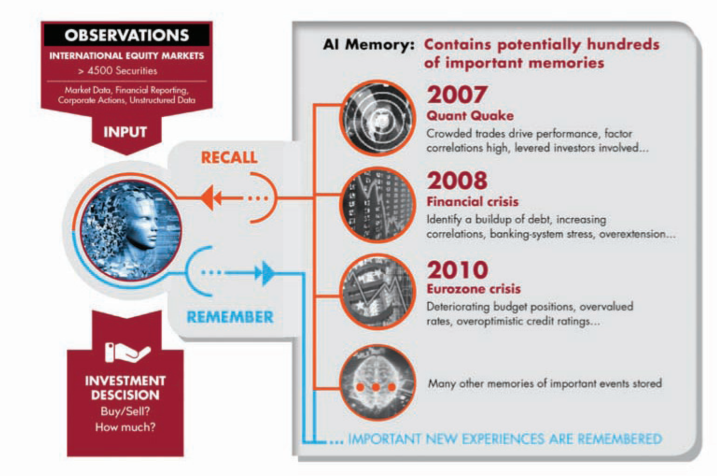
Source: Based on Lifelong machine learning, Chen et al. 2018

CONTINUAL LEARNING AUGMENTATION

A team I belong to from City University of London (Philps, Weyde, Garcez and Batchelor, 2018) developed an end-to-end learning system that acquires knowledge of different states (regimes) from multiple time series, and then applies this to a forecasting process that guides investment decisions. We call it *continual learning augmentation* (CLA) and it is applied as an open-world approach, belonging to a class of deep ML algorithms called memory-augmented neural nets (MANNs).

A base learner is chosen to drive CLA—for example, linear regression or a multilayered perceptron. This selected approach is run in a conventional way, stepping forward through time, forecasting time steps ahead. We then add a memory structure to this base learner. CLA's memory structure is designed to contain

Figure 3. A CLA System to Guide Investment Decisions


OBSERVATIONS
INTERNATIONAL EQUITY MARKETS
> 4500 Securities
Market Data, Financial Reporting, Corporate Actions, Unstructured Data
INPUT
RECALL
REMEMBER
INVESTMENT DESCISION
Buy/Sell?
How much?
AI Memory: Contains potentially hundreds of important memories
2007
Quant Quake
Crowded trades drive performance, factor correlations high, levered investors involved...
2008
Financial crisis
Identify a buildup of debt, increasing correlations, banking-system stress, overextension...
2010
Eurozone crisis
Deteriorating budget positions, overvalued rates, overoptimistic credit ratings...
Many other memories of important events stored
... IMPORTANT NEW EXPERIENCES ARE REMEMBERED

Source: Philps and colleagues, 2019

its base learner's parameters, which are remembered and recalled to improve future forecasting. Two observations allow us to develop this approach. First, we found it is possible to remember the most effective base learner parameterizations (model memories) over time, as patterns in the input data changed. Secondly, as **Figure 3** shows, we found it is possible to recall these model memories at a future time by recognising reoccurring patterns in the input data.

We tested the system in a trading simulation using multivariate time-series data from recent financial history, including the period leading up to the subprime crisis, the "quant quake," the post-quantitative easing (QE) era, and the (first) eurozone crisis. Base-learner parameterizations that appeared to best identify good (and bad) investments during these periods were stored as model memories that could be recalled when current events seemed to echo the past. For example, the approach recalled the QE-driven recovery in 2009 and identified this knowledge as the most pertinent to apply in stock-selection decisions during the stimulus-driven stock market rally in China in 2017.

We found the system would have significantly outperformed the investment returns of the simple, unaugmented base learners we tested in a global equities investment simulation between 2003 and 2017. We believe these returns would have put a CLA-driven investment strategy in the top 25% of managed funds by return over the study period.

Although CLA does not overcome the stability-plasticity dilemma, we have shown that CL can be effectively applied to specific, complex, real-world tasks.

CONCLUSION

In spite of its perceived complexity, end-to-end machine learning is likely to become an indispensable tool for forecasters. It will reduce human involvement in model development and, in doing, make outcomes more objective. Additionally, the complexity of ML approaches is a price worth paying if the result is richer information and automated learning.

This article has addressed the potential benefits of two advances in machine learning: AutoML and continual learning (CL). While AutoML is now a reality, offering forecasters a powerful, automated approach, the next generation of ML promises to be more powerful still.

CL will allow machines to learn over time, enabling generalizations across many tasks. While key research questions behind CL remain unanswered, this has not stopped the successful development of applied CL, of which continual learning augmentation (CLA) is an important example. So the seemingly simple question as to whether we "go simple or go complex" is not as simple as it seems.

REFERENCES

Batchelor, R. (2005), A Primer on Forecasting with Neural Networks, *Foresight*, Issue 2 (October), 37-43.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*, California: Stanford University Press.

Feurer, M., Klein, A, Eggensperger, K, Springenberg, J. B. & Hutter, F (2015). Efficient and Robust Automated Machine Learning, *Advances in Neural Information Processing Systems* 28 (NIPS 2015): 2962–2970.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Parwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A., Hermann, K.M., YoriZwols, Georg Ostrovski, Y., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K. & Hassabis, D. (2016). Hybrid Computing Using a Neural Network with Dynamic External Memory, *Nature*, volume 538 (27 October), 471–476.

Green, K.C. &. Armstrong, S. (2015). Simple Versus Complex Forecasting: The Evidence, *Journal of Business Research*, 68, 1678–1685.

Hergovich, A., Schott, R. & Burger, C. (2010). Biased Evaluation of Abstracts Depending on Topic and Conclusion: Further Evidence of a Confirmation Bias Within Scientific Psychology, *Current Psychology*, 29 (3): 188–209, doi:10.1007/s12144-010-9087-5

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory, *Neural Computation*, Volume 9, Issue 8 (November 15), 1735-1780.

Kotthoff, L., Thornton, C., Hoos, H. & Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA, *Journal of Machine Learning Research*, 18, 1-5.

Mermillod, M., Bugaiska, A. & Bonin, P. (2013). The Stability-Plasticity Dilemma: Investigating the Continuum from Catastrophic Forgetting to Age-Limited Learning Effects, *Frontiers in Psychology*, 4: 504.

Philps, D., Weyde, T. d'AvilaGarcez, A. & Batchelor, R . (2018). Continual Learning Augmented Investment Decisions, ***https://arxiv.org/abs/1812.02340***

Silver, D. (2015). Consolidation Using Sweep Task Rehearsal: Overcoming the Stability-Plasticity Problem, *Canadian AI 2015: Advances in Artificial Intelligence*, 307-322.

Tversky, A. & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability, *Cognitive Psychology*, 5 (2): 207–232. doi:10.1016/0010-0285(73)90033-9. ISSN 0010-0285.

Daniel Philps is head of Rothko Investment Strategies, an AI-driven investment group, and is an artificial intelligence researcher at City University of London.

Dan.philps@rothko.com

Acknowledgement. The author thanks Peter Simon, Data Scientist at DataRobot for his advice and fact checking on AutoML and to Professor Roy Batchelor of the City University of London for inspiring this entire endeavor.