PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
Deep learning, time series modelling

Keywords:
Deep neural networks, time series forecasting, uncertainty estimation, hybrid models, interpretability, counterfactual prediction

Author for correspondence:
Bryan Lim
e-mail: blim@robots.ox.ac.uk

arXiv:2004.13408v1 [stat.ML] 28 Apr 2020

THE ROYAL SOCIETY PUBLISHING

Time Series Forecasting With Deep Learning: A Survey

Bryan Lim[1] and Stefan Zohren[1]

[1]Department of Engineering Science, University of Oxford, Oxford, UK

Numerous deep learning architectures have been developed to accommodate the diversity of time series datasets across different domains. In this article, we survey common encoder and decoder designs used in both one-step-ahead and multi-horizon time series forecasting – describing how temporal information is incorporated into predictions by each model. Next, we highlight recent developments in hybrid deep learning models, which combine well-studied statistical models with neural network components to improve pure methods in either category. Lastly, we outline some ways in which deep learning can also facilitate decision support with time series data.

1. Introduction

Time series modelling has historically been a key area of academic research – forming an integral part of applications in topics such as climate modelling [1], biological sciences [2] and medicine [3], as well as commercial decision making in retail [4] and finance [5] to name a few. While traditional methods have focused on parametric models informed by domain expertise – such as autoregressive (AR) [6], exponential smoothing [7] or structural time series models [8] – modern machine learning methods provide a means to learn temporal dynamics in a purely data-driven manner [9]. With the increasing data availability and computing power in recent times, machine learning has become a vital part of the next generation of time series forecasting models.

Deep learning in particular has gained popularity in recent times, inspired by notable achievements in image classification [10], natural language processing [11] and reinforcement learning [12]. By incorporating bespoke architectural assumptions – or inductive biases [13] – that reflect the nuances of underlying datasets, deep neural networks are able to learn complex data representations [14], which alleviates the need for manual feature engineering and model design. The availability of open-source backpropagation frameworks [15, 16] has also simplified the network training, allowing for the customisation for network components and loss functions.

© The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Given the diversity of time-series problems across various domains, numerous neural network design choices have emerged. In this article, we summarise the common approaches to time series prediction using deep neural networks. Firstly, we describe the state-of-the-art techniques available for common forecasting problems – such as multi-horizon forecasting and uncertainty estimation. Secondly, we analyse the emergence of a new trend in hybrid models, which combine both domain-specific quantitative models with deep learning components to improve forecasting performance. Next, we outline two key approaches in which neural networks can be used to facilitate decision support, specifically through methods in interpretability and counterfactual prediction. Finally, we conclude with some promising future research directions in deep learning for time series prediction – specifically in the form of continuous-time and hierarchical models.

2. Deep Learning Architectures for Time Series Forecasting

Time series forecasting models predict future values of a target $y_{i,t}$ for a given entity i at time t. Each entity represents a logical grouping of temporal information – such as measurements from different weather stations in climatology, or vital signs from different patients in medicine – and can be observed at the same time. In the simplest case, one-step-ahead forecasting models take the form:

$$\hat{y}_{i,t+1} = f(\boldsymbol{y}_{i,t-k:t}, \boldsymbol{x}_{i,t-k:t}, \boldsymbol{s}_i), \tag{2.1}$$

where $\hat{y}_{i,t+1}$ is the model forecast, $\boldsymbol{y}_{i,t-k:t} = \{y_{i,t-k}, \ldots, y_{i,t}\}$, $\boldsymbol{x}_{i,t-k:t} = \{\boldsymbol{x}_{i,t-k}, \ldots, \boldsymbol{x}_{i,t}\}$ are observations of the target and exogenous inputs respectively over a look-back window k, s_i is static metadata associated with the entity (e.g. sensor location), and $f(.)$ is the prediction function learnt by the model. While we focus on univariate forecasting in this survey (i.e. 1-D targets), we note that the same components can be extended to multivariate models without loss of generality [17, 18, 19, 20, 21]. For notational simplicity, we omit the entity index i in subsequent sections unless explicitly required.

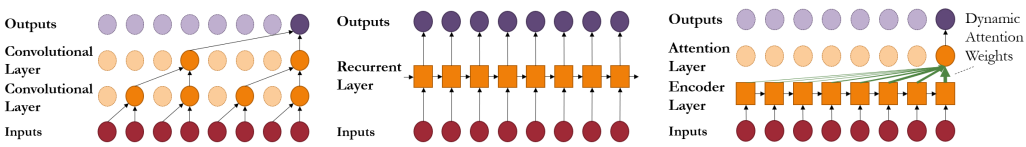
(a) Basic Building Blocks

Deep neural networks learn predictive relationships by using a series of non-linear layers to construct intermediate feature representations [14]. In time series settings, this can be viewed as encoding relevant historical information into a latent variable $\boldsymbol{z}_t$, with the final forecast produced using $\boldsymbol{z}_t$ alone:

$$f(\boldsymbol{y}_{t-k:t}, \boldsymbol{x}_{t-k:t+1}, \boldsymbol{s}) = g_{\text{dec}}(\boldsymbol{z}_t), \tag{2.2}$$

$$\boldsymbol{z}_t = g_{\text{enc}}(\boldsymbol{y}_{t-k:t}, \boldsymbol{x}_{t-k:t}, \boldsymbol{s}), \tag{2.3}$$

where $g_{\text{enc}}(.)$, $g_{\text{dec}}(.)$ are encoder and decoder functions respectively. These encoders and decoders hence form the basic building blocks of deep learning architectures, with the choice of network determining the types of relationships that can be learnt by our model. In this section, we examine modern design choices for encoders, as overviewed in Figure 1, and their relationship to traditional temporal models. In addition, we explore common network outputs and loss functions used in time series forecasting applications.



(a) CNN Model. (b) RNN Model. (c) Attention-based Model.

Figure 1: Incorporating temporal information using different encoder architectures.

(i) Convolutional Neural Networks

Traditionally designed for image datasets, convolutional neural networks (CNNs) extract local relationships that are invariant across spatial dimensions [10, 22]. To adapt CNNs to time series datasets, researchers utilise multiple layers of causal convolutions [23, 24, 25] – i.e. convolutional filters designed to ensure only past information is used for forecasting. For an intermediate feature at hidden layer l, each causal convolutional filter take the form below:

$$\boldsymbol{h}_t^{l+1} = A\Big(\left(\boldsymbol{W} * \boldsymbol{h}\right)(l,t) \Big), \tag{2.4}$$

$$\left(\boldsymbol{W} * \boldsymbol{h}\right)(l,t) = \sum_{\tau=0}^{k} \boldsymbol{W}(l,\tau)\boldsymbol{h}_{t-\tau}^{l}, \tag{2.5}$$

where $\boldsymbol{h}_t^l \in \mathbb{R}^{\mathcal{H}_{in}}$ is an intermediate state at layer l at time t, $*$ is the convolution operator, $\boldsymbol{W}(l,\tau) \in \mathbb{R}^{\mathcal{H}_{out}\times\mathcal{H}_{in}}$ is a fixed filter weight at layer l, and $A(.)$ is an activation function, such as a sigmoid function, representing any architecture-specific non-linear processing.

Considering the 1-D case, we can see that Equation (2.5) bears a strong resemblance to finite impulse response (FIR) filters in digital signal processing [26]. This leads to two key implications for temporal relationships learnt by CNNs. Firstly, in line with the spatial invariance assumptions for standard CNNs, temporal CNNs assume that relationships are time-invariant – using the same set of filter weights at each time step and across all time. In addition, CNNs are only able to use inputs within its defined lookback window, or receptive field, to make forecasts. As such, the receptive field size k needs to be tuned carefully to ensure that the model can make use of all relevant historical information. It is worth noting that a single causal CNN layer is equivalent to an auto-regressive (AR) model.

Dilated Convolutions Using standard convolutional layers can be computational challenging where long-term dependencies are significant, as the number of parameters scales directly with the size of the receptive field. To alleviate this, modern architectures frequently make use of dilated covolutional layers [23, 24], which extend Equation (2.5) as below:

$$\left(\boldsymbol{W} * \boldsymbol{h}\right)(l,t,d_l) = \sum_{\tau=0}^{\lfloor k/d_l \rfloor} \boldsymbol{W}(l,\tau)\boldsymbol{h}_{t-d_l\tau}^{l}, \tag{2.6}$$

where $\lfloor . \rfloor$ is the floor operator and d_l is a layer-specific dilation rate. Dilated convolutions can hence be interpreted as convolutions of a down-sampled version of the lower layer features – reducing resolution to incorporate information from the distant past. As such, by increasing the dilation rate with each layer, dilated convolutions can gradually aggregate information at different time blocks, allowing for more history to be used in an efficient manner. With the WaveNet architecture of [23] for instance, dilation rates are increased in powers of 2 with adjacent time blocks aggregated in each layer – allowing for 2^l time steps to be used at layer l as shown in Figure 1a.

(ii) Recurrent Neural Networks

Recurrent neural networks (RNNs) have had a historically been used in sequence modelling [22], with strong results on a variety of natural language processing tasks [27]. Given the natural interpretation of time series data as sequences of inputs and targets, many RNN-based architectures have been developed for temporal forecasting applications [28, 29, 30, 31]. At its core, RNN cells contain an internal memory state which acts as a compact summary of past information. The memory state is recursively updated using with new observations at each time step as shown in Figure 1b, i.e.:

$$\boldsymbol{z}_t = \nu\left(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{x}}_t\right), \tag{2.7}$$

Where $\boldsymbol{z}_t \in \mathbb{R}^{\mathcal{H}}$ here is the hidden internal state of the RNN, $\tilde{\boldsymbol{x}}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{s}]$ are the combined inputs into the RNN at each step, and $\nu(.)$ is the learnt memory update function. For instance, the Elman RNN [32], one of the simplest RNN variants, would take the form below:

$$y_{t+1} = \gamma_y(\boldsymbol{W}_y \boldsymbol{z}_t + \boldsymbol{b}_y), \tag{2.8}$$

$$\boldsymbol{z}_t = \gamma_z(\boldsymbol{W}_{z_1} \boldsymbol{z}_{t-1} + \boldsymbol{W}_{z_2} \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_z), \tag{2.9}$$

Where $\boldsymbol{W}_.$, $\boldsymbol{b}_.$ are the linear weights and biases of the network respectively, and $\gamma_y(.), \gamma_z(.)$ are network activation functions. Note that RNNs do not require the explicit specification of a lookback window as per the CNN case. From a signal processing perspective, the main recurrent layer – i.e. Equation (2.9) – thus resembles a non-linear version of infinite impulse response (IIR) filters.

Long Short-term Memory Due to the infinite lookback window, older variants of RNNs can suffer from limitations in learning long-range dependencies in the data [33] – due to issues with exploding and vanishing gradients [22]. Intuitively this can be seen as a form of resonance in the memory state. Long Short-Term Memory networks (LSTMs) [34] were hence developed to address these limitations, by improving gradient flow within the network. This is achieved through the use of a cell state $\boldsymbol{c}_t$ which stores long-term information, modulated through a series of gates as below:

$$\textit{Input gate}: \quad \boldsymbol{i}_t = \sigma(\boldsymbol{W}_{i_1} \boldsymbol{h}_{t-1} + \boldsymbol{W}_{i_2} \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_i), \tag{2.10}$$

$$\textit{Output gate}: \quad \boldsymbol{o}_t = \sigma(\boldsymbol{W}_{o_1} \boldsymbol{h}_{t-1} + \boldsymbol{W}_{o_2} \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_o), \tag{2.11}$$

$$\textit{Forget gate}: \quad \boldsymbol{f}_t = \sigma(\boldsymbol{W}_{f_1} \boldsymbol{h}_{t-1} + \boldsymbol{W}_{f_2} \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_f), \tag{2.12}$$

where $\boldsymbol{h}_{t-1}$ is the hidden state of the LSTM, and $\sigma(.)$ is the sigmoid activation function. The gates modify the hidden and cell states of the LSTM as below:

$$\textit{Hidden state}: \quad \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), \tag{2.13}$$

$$\textit{Cell state}: \quad \boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_{c_1} \boldsymbol{h}_{t-1} + \boldsymbol{W}_{c_2} \tilde{\boldsymbol{x}}_t + \boldsymbol{b}_c), \tag{2.14}$$

Where $\odot$ is the element-wise (Hadamard) product, and $\tanh(.)$ is the tanh activation function.

Relationship to Bayesian Filtering As examined in [30], Bayesian filters [35] and RNNs are both similar in their maintenance of a hidden state which is recursively updated over time. For Bayesian filters, such as the Kalman filter [36], inference is performed by updating the sufficient statistics of the latent state – using a series of state transition and error correction steps. As the Bayesian filtering steps use deterministic equations to modify sufficient statistics, the RNN can be viewed as a simultaneous approximation of both steps – with the memory vector containing all relevant information required for prediction.

(iii) Attention Mechanisms

The development of attention mechanisms [37, 38] has also lead to improvements in long-term dependency learning – with Transformer architectures achieving state-of-the-art performance in multiple natural language processing applications [11, 39, 40]. Attention layers aggregate temporal features using dynamically generated weights (see Figure 1c), allowing the network to directly focus on significant time steps in the past – even if they are very far back in the lookback window. In general, attention weighting takes the form below:

$$\tilde{\boldsymbol{h}}_t = \sum_{\tau=0}^{k} \alpha(t, \tau) \boldsymbol{h}_{t-\tau}, \tag{2.15}$$

Where $\boldsymbol{h}_{t-\tau}$ is an intermediate feature extracted by the lower levels of the network, $\alpha(t, \tau) \in [0, 1]$ is the attention weight for $t - \tau$ generated at time t, and $\tilde{\boldsymbol{h}}_t$ is the context vector output of the attention layer.

Recent work has also demonstrated the benefits of using attention mechanisms in time series forecasting applications, with improved performance over comparable recurrent networks [41, 42, 43]. For instance, [41] use attention to aggregate features extracted by RNN encoders, with attention weights produced as below:

$$\boldsymbol{\alpha}(t) = \text{softmax}(\boldsymbol{\eta}_t), \tag{2.16}$$

$$\boldsymbol{\eta}_t = \mathbf{W}_{\eta_1} \tanh(\mathbf{W}_{\eta_2} \tilde{\boldsymbol{h}}_{t-1} + \mathbf{W}_{\eta_3} \boldsymbol{h}_t + \boldsymbol{b}_\eta), \tag{2.17}$$

where $\boldsymbol{\alpha}(t) = [\alpha(t,0), \dots \alpha(t,k)]$ is a vector of attention weights, $\tilde{\boldsymbol{h}}_{t-1}, \boldsymbol{h}_t$ are outputs from LSTM encoders used for feature extraction, and softmax(.) is the softmax activation function. More recently, Transformer architectures have also been considered in [42, 43], which apply scalar-dot product self-attention [39] to features extracted within the lookback window. From a time series modelling perspective, attention provides two key benefits. Firstly, networks with attention are able to directly attend to any significant events that occur. In retail forecasting applications, for example, this includes holiday or promotional periods which can have a positive effect on sales. Secondly, as shown in [43], attention-based networks can also learn regime-specific temporal dynamics – by using distinct attention weight patterns for each regime.

(iv) Outputs and Loss Functions

Given the flexibility of neural networks, deep neural networks have been used to model both discrete [44] and continuous [28, 45] targets – by customising of decoder and output layer of the neural network to match the desired target type. In one-step-ahead prediction problems, this can be as simple as combining a linear transformation of encoder outputs (i.e. Equation (2.2)) together with an appropriate output activation for the target. Regardless of the form of the target, predictions can be further divided into two different categories – point estimates and probabilistic forecasts.

Point Estimates A common approach to forecasting is to determine the expected value of a future target. This essentially involves reformulating the problem to a classification task for discrete outputs (e.g. forecasting future events), and regression task for continuous outputs – using the encoders described above. For the binary classification case, the final layer of the decoder then features a linear layer with a sigmoid activation function – allowing the network to predict the probability of event occurrence at a given time step. For one-step-ahead forecasts of binary and continuous targets, networks are trained using binary cross-entropy and mean square error loss functions respectively:

$$\mathcal{L}_{classification} = -\frac{1}{T} \sum_{t=1}^{T} y_t \log(\hat{y}_t) + (1 - y_t) \log(1 - \hat{y}_t) \tag{2.18}$$

$$\mathcal{L}_{regression} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \tag{2.19}$$

Probabilistic Outputs While point estimates are crucial to predicting the future value of a target, understanding the uncertainty of a model's forecast can be useful for decision makers in different domains. When forecast uncertainties are wide, for instance, model users can exercise more caution when incorporating predictions into their decision making, or alternatively rely on other sources of information. In some applications, such as financial risk management, having access to the full predictive distribution will allow decision makers to optimise their actions in the presence of rare events – e.g. allowing risk managers to insulate portfolios against market crashes.

A common way to model uncertainties is to use deep neural networks to generate parameters of known distributions [18, 28, 29]. For example, Gaussian distributions are typically used for forecasting problems with continuous targets, with the networks outputting means and variance

parameters for the predictive distributions at each step as below:

$$y_{t+\tau} \sim N(\mu(t,\tau), \zeta(t,\tau)^2), \tag{2.20}$$

$$\mu(t,\tau) = \boldsymbol{W}_\mu \boldsymbol{h}_t^L + \boldsymbol{b}_\mu, \tag{2.21}$$

$$\zeta(t,\tau) = \text{softplus}(\boldsymbol{W}_\Sigma \boldsymbol{h}_t^L + \boldsymbol{b}_\Sigma), \tag{2.22}$$

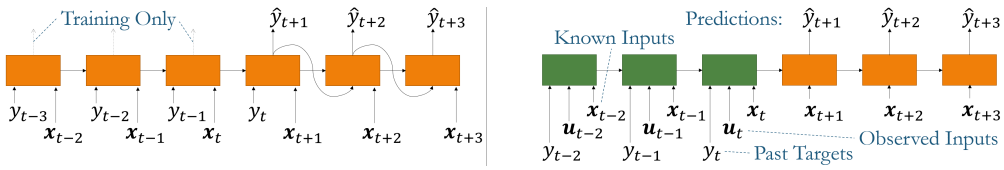
where $\boldsymbol{h}_t^L$ is the final layer of the network, and softplus(.) is the softplus activation function to ensure that standard deviations take only positive values.

(b) Multi-horizon Forecasting Models

In many applications, it is often beneficial to have access to predictive estimates at multiple points in the future – allowing decision makers to visualise trends over a future horizon, and optimise their actions across the entire path. From a statistical perspective, multi-horizon forecasting can be viewed as a slight modification of one-step-ahead prediction problem (i.e. Equation (2.1)) as below:

$$\hat{y}_{t+\tau} = f(\boldsymbol{y}_{t-k:t}, \boldsymbol{x}_{t-k:t}, \boldsymbol{u}_{t-k:t+h}, \boldsymbol{s}, \tau), \tag{2.23}$$

where $\tau \in \{1, \ldots, \tau_{max}\}$ is a discrete forecast horizon, $\boldsymbol{x}_t$ are known future inputs (e.g. date information, such as the day-of-week or month) across the entire horizon, and $\boldsymbol{x}_t$ are inputs that can only be observed historically. In line with traditional econometric approaches [46, 47], deep learning architectures for multi-horizon forecasting can be divided into iterative and direct methods – as shown in Figure 2 and described in detail below.



(a) Iterative Methods (b) Direct Methods

Figure 2: Main types of multi-horizon forecasting models.

(i) Iterative Methods

Iterative approaches to multi-horizon forecasting typically make use of autoregressive deep learning architectures [28, 30, 31, 42] – producing multi-horizon forecasts by recursively feeding samples of the target into future time steps (see Figure 2a). By repeating the procedure to generate multiple trajectories, forecasts are then produced using the sampling distributions for target values at each step. For instance, predictive means can be obtained using the Monte Carlo estimate $\hat{\boldsymbol{y}}_{t+\tau} = \sum_{j=1}^{J} \tilde{\boldsymbol{y}}_{t+\tau}^{(j)}/J$, where $\tilde{\boldsymbol{y}}_{t+\tau}^{(j)}$ is a sample taken based on the model of Equation (2.20). As autoregressive models are trained in the exact same fashion as one-step-ahead prediction models, the iterative approach allows for the easy generalisation of standard models to multi-step forecasting. However, the recursive structure of iterative methods assumes that all inputs but the target are known at run-time – requiring only samples of the target to be fed into future time steps. This can be a limitation in many practical scenarios where observed inputs exist, motivating the need for more flexible methods.

(ii) Direct Methods

Direct methods alleviate the issues with iterative methods by producing forecasts directly using all available inputs. They typically make use of sequence-to-sequence architectures [41, 43, 45], using an encoder to summarise past information (i.e. targets, observed inputs and a priori known

inputs), and a decoder to combine them with known future inputs – as depicted in Figure 2b. While avoiding the need for recursion, direct methods require the specification of a maximum forecast horizon (i.e. τ_{max}), with predictions made only at predefined discrete intervals.

3. Incorporating Domain Knowledge with Hybrid Models

Despite its popularity, the efficacy of machine learning for time series prediction has historically been questioned – as evidenced by forecasting competitions such as the M-competitions [48]. Prior to the M4 competition of 2018 [49], the prevailing wisdom was that sophisticated methods do not produce more accurate forecasts, and simple models with ensembling had a tendency to do better [50, 51, 52]. Two key reasons have been identified to explain the underperformance of machine methods. Firstly, the flexibility of machine learning methods can be a double-edged sword – making them prone to overfitting [52]. Secondly, similar to stationarity requirements of statistical models, machine learning models can be sensitive to how inputs are pre-processed [17, 28, 52], which ensure that data distributions at training and test time are similar.

A recent trend in deep learning has been in developing hybrid models which address these limitations, demonstrating improved performance over pure statistical or machine learning models in a variety of applications [29, 53, 54, 55]. Hybrid methods combine well-studied quantitative time series models together with deep learning – using deep neural networks to generate model parameters at each time step. On the one hand, hybrid models allow domain experts to inform neural network training using prior information – reducing the hypothesis space of the network and improving generalisation. This is especially useful for small datasets [29], where there is a greater risk of overfitting for deep learning models. Furthermore, hybrid models allow for the separation of stationary and non-stationary components, and avoid the need for custom input pre-processing. An example of this is the Exponential Smoothing RNN (ES-RNN) [53], winner of the M4 competition, which uses exponential smoothing to capture non-stationary trends and learns additional effects with the RNN. In general, hybrid models utilise deep neural networks in two manners: a) to encode time-varying parameters for non-probabilistic parametric models [53, 54, 56], and b) to produce distribution parameters for probabilistic models [29, 31, 55].

(a) Non-probabilistic Hybrid Models

With parametric time series models, forecasting equations are typically defined analytically and provide point forecasts for future targets. Non-probabilistic hybrid models hence modify these forecasting equations to combine statistical and deep learning components. The ES-RNN for example, utilises the update equations of the Holt-Winters exponential smoothing model [57] – combining multiplicative level and seasonality components with deep learning outputs as below:

$$\hat{y}_{i,t+\tau} = \exp(\boldsymbol{W}_{ES}\boldsymbol{h}^{L}_{i,t+\tau} + \boldsymbol{b}_{ES}) \times l_{i,t} \times \gamma_{i,t+\tau}, \tag{3.1}$$

$$l_{i,t} = \beta_1^{(i)} y_{i,t}/\gamma_{i,t} + (1-\beta_1^{(i)}) l_{i,t-1}, \tag{3.2}$$

$$\gamma_{i,t} = \beta_2^{(i)} y_{i,t}/l_{i,t} + (1-\beta_2^{(i)}) \gamma_{i,t-\kappa}, \tag{3.3}$$

where $\boldsymbol{h}^{L}_{i,t+\tau}$ is the final layer of the network for the τth-step-ahead forecast, $l_{i,t}$ is a level component, $\gamma_{i,t}$ is a seasonality component with period κ, and $\beta_1^{(i)}, \beta_2^{(i)}$ are entity-specific static coefficients. From the above equations, we can see that the exponential smoothing components $(l_{i,t}, \gamma_{i,t})$ handle the broader (e.g. exponential) trends within the datasets, reducing the need for additional input scaling.

(b) Probabilistic Hybrid Models

Probabilistic hybrid models can also be used in applications where distribution modelling is important – utilising probabilistic generative models for temporal dynamics such as Gaussian

processes [31] and linear state space models [29]. Rather than modifying forecasting equations, probabilistic hybrid models use neural networks produce parameters for predictive distributions at each step. For instance, Deep State Space Models [29] encode time-varying parameters for linear state space models as below – performing inference via the Kalman filtering equations [36]:

$$y_t = \boldsymbol{a}(\boldsymbol{h}^L_{i,t+\tau})^T \boldsymbol{l}_t + \phi(\boldsymbol{h}^L_{i,t+\tau})\epsilon_t, \tag{3.4}$$

$$\boldsymbol{l}_t = \boldsymbol{F}(\boldsymbol{h}^L_{i,t+\tau})\boldsymbol{l}_{t-1} + \boldsymbol{q}(\boldsymbol{h}^L_{i,t+\tau}) + \boldsymbol{\Sigma}(\boldsymbol{h}^L_{i,t+\tau}), \tag{3.5}$$

where $\boldsymbol{l}_t$ is the hidden latent state, $\boldsymbol{a}(.)$, $\boldsymbol{F}(.)$, $\boldsymbol{q}(.)$ are linear transformations of $\boldsymbol{h}^L_{i,t+\tau}$, $\phi(.)$, $\boldsymbol{\Sigma}(.)$ are linear transformations with softmax activations, and $\epsilon_t \sim N(0,1)$, $\boldsymbol{\Sigma}_t \sim N(0, \mathbb{I})$ are standard normal random variables.

4. Facilitating Decision Support Using Deep Neural Networks

Although model builders are mainly concerned with the accuracy of their forecasts, end-users typically use predictions to *guide their future actions*. For instance, doctors can make use of clinical forecasts (e.g. probabilities of disease onset and mortality) to help them prioritise tests to order, formulate a diagnosis and determine a course of treatment. As such, while time series forecasting is a crucial preliminary step, a better understanding of both temporal dynamics and the motivations behind a model's forecast can help users further optimise their actions. In this section, we explore two directions in which neural networks have been extended to facilitate decision support with time series data – focusing on methods in interpretability and causal inference.

(a) Interpretability With Time Series Data

With the deployment of neural networks in mission-critical applications [58], there is a increasing need to understand both *how* and *why* a model makes a certain prediction. Moreover, end-users can have little prior knowledge with regards to the relationships present in their data, with datasets growing in size and complexity in recent times. Given the black-box nature of standard neural network architectures, a new body of research has emerged in methods for interpreting deep learning models, which we present in depth below.

Techniques for Post-hoc Interpretability Post-hoc interpretable models are developed to interpret trained networks, and helping to identify important features or examples without modifying the original weights. Methods can mainly be divided into two main categories. Firstly, one possible approach is to apply simpler interpretable surrogate models between the inputs and outputs of the neural network, and rely on the approximate model to provide explanations. For instance, Local Interpretable Model-Agnostic Explanations (LIME) [59] identify relevant features by fitting instance-specific linear models to perturbations of the input, with the linear coefficients providing a measure of importance. Shapley additive explanations (SHAP) [60] provide another surrogate approach, which utilises Shapley values from cooperative game theory to identify important features across the dataset. Next, gradient-based method – such as saliency maps [61, 62] and influence functions [63] – have been proposed, which analyse network gradients to determine which input features have the greatest impact on loss functions. While post-hoc interpretability methods can help with feature attributions, they typically ignore any sequence dependencies between inputs – making it difficult to apply them to complex time series datasets.

Inherent Interpretability with Attention Weights An alternative approach is to directly design architectures with explainable components, typically in the form of strategically placed attention layers. As attention weights are produced as outputs from a softmax layer, the weights are constrained to sum to 1, i.e. $\sum_{\tau=0}^{k} \alpha(t,\tau) = 1$. For time series models, the outputs of Equation (2.15) can hence also be interpreted as a weighted average over temporal features, using the weights supplied by the attention layer at each step. An analysis of attention weights can then be used to

understand the relative importance of features at each time step. Instance-wise interpretability studies have been performed in [42, 44, 64], where the authors used specific examples to show how the magnitudes of $\alpha(t, \tau)$ can indicate which time points were most significant for predictions. By analysing distributions of attention vectors across time, [43] also shows how attention mechanisms can be used to identify persistent temporal relationships – such as seasonal patterns – in the dataset.

(b) Counterfactual Predictions & Causal Inference Over Time

In addition to understanding the relationships learnt by the networks, deep learning can also help to facilitate decision support by producing predictions outside of their observational datasets, or counterfactual forecasts. Counterfactual predictions are particularly useful for scenario analysis applications – allowing users to evaluate how different sets of actions can impact target trajectories. This can be useful both from a historical angle, i.e. determining what would have happened if a different set of circumstances had occurred, and from a forecasting perspective, i.e. determining which actions to take to optimise future outcomes.

While a large class of deep learning methods exists for estimating causal effects in static settings [65, 66, 67], the key challenge in time series datasets is the presence of time-dependent confounding effects. This arises due to circular dependencies when actions that can affect the target are also conditional on observations of the target. Without any adjusting for time-dependent confounders, straightforward estimations techniques can results in biased results, as shown in [68]. Recently, several methods have emerged to train deep neural networks while adjusting for time-dependent confounding, based on extensions of statistical techniques and the design of new loss functions. With statistical methods, [69] extend the inverse-probability-of-treatment-weighting (IPTW) approach of marginal structural models in epidemiology – using one set of networks to estimate treatment application probabilities, and a sequence-to-sequence model to learn unbiased predictions. Another approach in [70] extends the G-computation framework, jointly modelling distributions of the target and actions using deep learning. In addition, new loss functions have have been proposed in [71], which adopts domain adversarial training to learn balanced representations of patient history.

5. Conclusions and Future Directions

With the growth in data availability and computing power in recent times, deep neural networks architectures have achieved much success in forecasting problems across multiple domains. In this article, we survey the main architectures used for time series forecasting – highlighting the key building blocks used in neural network design. We examine how they incorporate temporal information for one-step-ahead predictions, and describe how they can be extended for use in multi-horizon forecasting. Furthermore, we outline the recent trend of hybrid deep learning models, which combine statistical and deep learning components to outperform pure methods in either category. Finally, we summarise two ways in which deep learning can be extended to improve decision support over time, focusing on methods in interpretability and counterfactual prediction.

Although a large number of deep learning models have been developed for time series forecasting, some limitations still exist. Firstly, deep neural networks typically require time series to be discretised at regular intervals, making it difficult to forecast datasets where observations can be missing or arrive at random intervals. While some preliminary research on continuous-time models has been done via Neural Ordinary Differential Equations [72], additional work needs to be done to extend this work for datasets with complex inputs (e.g. static variables) and to benchmark them against existing models. In addition, as mentioned in [73], time series often have a hierarchical structure with logical groupings between trajectories – e.g. in retail forecasting, where product sales in the same geography can be affected by common trends. As such, the development of architectures which explicit account for such hierarchies could be an interesting research direction, and potentially improve forecasting performance over existing univariate or multivariate models.

Competing Interests. The author(s) declare that they have no competing interests.

References

1 Mudelsee M. Trend analysis of climate time series: A review of methods. Earth-Science Reviews. 2019;190:310 – 322.
2 Stoffer DS, Ombao H. Editorial: Special issue on time series analysis in the biological sciences. Journal of Time Series Analysis. 2012;33(5):701–703.
3 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019 Jan;25(1):44–56.
4 Böse JH, Flunkert V, Gasthaus J, Januschowski T, Lange D, Salinas D, et al. Probabilistic Demand Forecasting at Scale. Proc VLDB Endow. 2017 Aug;10(12):1694–1705.
5 Andersen TG, Bollerslev T, Christoffersen PF, Diebold FX. Volatility Forecasting. National Bureau of Economic Research; 2005. 11188.
6 Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-Day; 1976.
7 Gardner Jr ES. Exponential smoothing: The state of the art. Journal of Forecasting. 1985;4(1):1–28.
8 Harvey AC. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press; 1990.
9 Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews. 2010;29(5-6):594–621.
10 Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25 (NIPS); 2012. p. 1097–1105.
11 Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 4171–4186.
12 Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529:484–503.
13 Baxter J. A Model of Inductive Bias Learning. J Artif Int Res. 2000;12(1):149âĂŞ198.
14 Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(8):1798–1828.
15 Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Software available from tensorflow.org. Available from: http://tensorflow.org/.
16 Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32; 2019. p. 8024–8035.
17 Sen R, Yu HF, , Dhillon I. Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting. In: Advances in Neural Information Processing Systems (NeurIPS); 2019. .
18 Wen R, Torkkola K. Deep Generative Quantile-Copula Models for Probabilistic Forecasting. In: ICML Time Series Workshop; 2019. .
19 Li Y, Yu R, Shahabi C, Liu Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In: (Proceedings of the International Conference on Learning Representations ICLR); 2018. .
20 Ghaderi A, Sanandaji BM, Ghaderi F. Deep Forecast: Deep Learning-based Spatio-Temporal Forecasting. In: ICML Time Series Workshop; 2017. .
21 Salinas D, Bohlke-Schneider M, Callot L, Medico R, Gasthaus J. High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes. In: Advances in Neural Information Processing Systems (NeurIPS); 2019. .
22 Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. http://www.deeplearningbook.org.
23 van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: A Generative Model for Raw Audio. arXiv e-prints. 2016 Sep;p. arXiv:1609.03499.
24 Bai S, Zico Kolter J, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv e-prints. 2018;p. arXiv:1803.01271.
25 Borovykh A, Bohte S, Oosterlee CW. Conditional Time Series Forecasting with Convolutional Neural Networks. arXiv e-prints. 2017;p. arXiv:1703.04691.

26 Lyons RG. Understanding Digital Signal Processing (2nd Edition). USA: Prentice Hall PTR; 2004.
27 Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine. 2018;13(3):55–75.
28 Salinas D, Flunkert V, Gasthaus J. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. arXiv e-prints. 2017;p. arXiv:1704.04110.
29 Rangapuram SS, Seeger MW, Gasthaus J, Stella L, Wang Y, Januschowski T. Deep State Space Models for Time Series Forecasting. In: Advances in Neural Information Processing Systems (NIPS); 2018. .
30 Lim B, Zohren S, Roberts S. Recurrent Neural Filters: Learning Independent Bayesian Filtering Steps for Time Series Prediction. In: International Joint Conference on Neural Networks (IJCNN); 2020. .
31 Wang Y, Smola A, Maddix D, Gasthaus J, Foster D, Januschowski T. Deep Factors for Forecasting. In: Proceedings of the International Conference on Machine Learning (ICML); 2019. .
32 Elman JL. Finding structure in time. Cognitive Science. 1990;14(2):179 – 211.
33 Kolen JF, Kremer SC. In: Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies; 2001. p. 237–243.
34 Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997 Nov;9(8):1735–1780.
35 Srkk S. Bayesian Filtering and Smoothing. Cambridge University Press; 2013.
36 Kalman RE. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering. 1960;82(1):35.
37 Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015. .
38 Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. .
39 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems (NIPS); 2017. .
40 Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL); 2019. .
41 Fan C, Zhang Y, Pan Y, Li X, Zhang C, Yuan R, et al. Multi-Horizon Time Series Forecasting with Temporal Attention Learning. In: Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD); 2019. .
42 Li S, Jin X, Xuan Y, Zhou X, Chen W, Wang YX, et al. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In: Advances in Neural Information Processing Systems (NeurIPS); 2019. .
43 Lim B, Arik SO, Loeff N, Pfister T. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv e-prints. 2019;p. arXiv:1912.09363.
44 Choi E, Bahadori MT, Sun J, Kulas JA, Schuetz A, Stewart WF. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In: Advances in Neural Information Processing Systems (NIPS); 2016. .
45 Wen R, et al. A Multi-Horizon Quantile Recurrent Forecaster. In: NIPS 2017 Time Series Workshop; 2017. .
46 Taieb SB, Sorjamaa A, Bontempi G. Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing. 2010;73(10):1950 – 1957.
47 Marcellino M, Stock J, Watson M. A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series. Journal of Econometrics. 2006;135:499–526.
48 Hyndman R. A brief history of forecasting competitions. International Journal of Forecasting. 2020;36(1):7–14.
49 The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting. 2020;36(1):54 – 74.
50 Fildes R, Hibon M, Makridakis S, Meade N. Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting. 1998;14(3):339 – 358.

51 Makridakis S, Hibon M. The M3-Competition: results, conclusions and implications. International Journal of Forecasting. 2000;16(4):451 – 476. The M3- Competition.

52 Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLOS ONE. 2018 03;13(3):1–26.

53 Smyl S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting. 2020;36(1):75 – 85. M4 Competition.

54 Lim B, Zohren S, Roberts S. Enhancing Time-Series Momentum Strategies Using Deep Neural Networks. The Journal of Financial Data Science. 2019;.

55 Grover A, Kapoor A, Horvitz E. A Deep Hybrid Model for Weather Forecasting. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD); 2015. .

56 Binkowski M, Marti G, Donnat P. Autoregressive Convolutional Neural Networks for Asynchronous Time Series. In: Proceedings of the International Conference on Machine Learning (ICML); 2018. .

57 Winters PR. Forecasting Sales by Exponentially Weighted Moving Averages. Management Science. 1960;6(3):324–342.

58 Moraffah R, Karami M, Guo R, Raglin A, Liu H. Causal Interpretability for Machine Learning – Problems, Methods and Evaluation. arXiv e-prints. 2020;p. arXiv:2003.03934.

59 Ribeio M, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In: KDD; 2016. .

60 Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NIPS); 2017. .

61 Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv e-prints. 2013;p. arXiv:1312.6034.

62 Siddiqui SA, Mercier D, Munir M, Dengel A, Ahmed S. TSViz: Demystification of Deep Learning Models for Time-Series Analysis. IEEE Access. 2019;7:67027–67040.

63 Koh PW, Liang P. Understanding Black-box Predictions via Influence Functions. In: Proceedings of the International Conference on Machine Learning(ICML; 2017. .

64 Bai T, Zhang S, Egleston BL, Vucetic S. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD); 2018. .

65 Yoon J, Jordon J, van der Schaar M. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In: International Conference on Learning Representations (ICLR); 2018. .

66 Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep IV: A Flexible Approach for Counterfactual Prediction. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. .

67 Alaa AM, Weisz M, van der Schaar M. Deep Counterfactual Networks with Propensity Dropout. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. .

68 Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359.

69 Lim B, Alaa A, van der Schaar M. Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks. In: NeurIPS; 2018. .

70 Li R, Shahn Z, Li J, Lu M, Chakraborty P, Sow D, et al. G-Net: A Deep Learning Approach to G-computation for Counterfactual Outcome Prediction Under Dynamic Treatment Regimes. arXiv e-prints. 2020;p. arXiv:2003.10551.

71 Bica I, Alaa AM, Jordon J, van der Schaar M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In: International Conference on Learning Representations(ICLR); 2020. .

72 Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. Neural Ordinary Differential Equations. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS); 2018. .

73 Fry C, Brundage M. The M4 Forecasting Competition – A Practitioner's View. International Journal of Forecasting. 2019;.