Stochastic Machine Learning
Chapter 03 - Deep time series

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

Deep time series modeling

- For an overview, see: Lim and Zohren (2021).
- In a time series setting for $X_0, X_2, X_2, \dots$ we generally have the following structure:

$$F_{X_t} = F(X_{t-1}, X_{t-2}, \dots, \epsilon_{t-1}, \epsilon_{t-2}, \dots)$$

  and probably some additional covariables. Here $\epsilon_0, \epsilon_1, \dots$ represents the noise random variables and those are typically i.i.d. and standardized
- The goal is typically a prediction, for example $\hat{X}_t$ of the future value, and the estimator is a function of the available data:

$$\hat{X}_t = f(X_{t-1}, X_{t-2}, \dots).$$

- Most neural networks aim at learning a feature representation - some hidden features which describe the data very well. If we denote the hidden features by Z, we obtain the **decoding mechanism**

$$\hat{X}_t = f_{\textbf{dec}}(Z_t)$$

  and the **encoding mechanism**

$$Z_t = f_{\textbf{enc}}(X_{t-1}, X_{t-2}, \dots).$$

- Recall that also the LSTM hat in the background a cell state - that would be a learned feature.

Deep probabilistic modelling

- More generally, we might want to predict the whole distribution of X_t and not only the value of X_t.
- Recall that a GARCH time series was of the type

$$\begin{cases} X_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \end{cases}, \tag{1}$$

- If we want to describe this more generally, we think of a family of distribution

$$(P_\theta : \theta \in \Theta)$$

and the mapping we are interested in is

$$P(X_t \in \cdot) = P_{F(X_{t-1}, X_{t-2}, \dots)}$$

- In deep probabilistic modeling one wants to learn the mapping F.
- Note that if $\Theta \subset \mathbb{R}^d$, our universal approximation results apply. In the more general setting not ...
- The simplest example is (like in (1) that

$$X_t \sim \mathcal{N}(\mu(X_{t-1}, X_{t-2}, \dots), \sigma(X_{t-1}, X_{t-2}, \dots))$$

and both μ and σ is represented through a neural network.

Further approaches for time series modelling

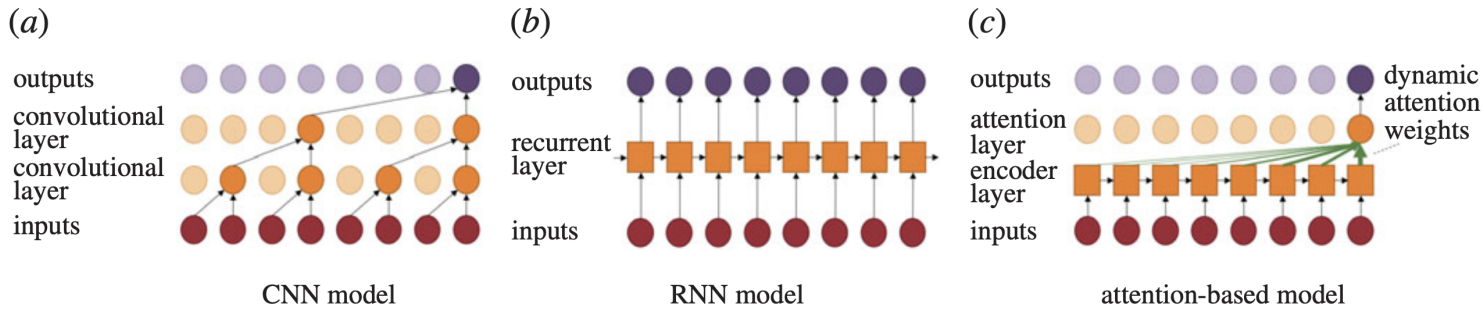

Figure 1. Incorporating temporal information using different encoder architectures. (*a*) CNN model, (*b*) RNN model and (*c*) attention-based model. (Online version in colour.)

Source: Lim and Zohren (2021)

- Also **attention mechnanisms** have proved useful which we shortly want to revisit.
- The idea is that to each time point t intermediate features $V_{t-1}, V_{t-2}, \ldots$ are produce by the neural network which are then weighted by the attention layer
- This layer is of the form

$$h_t = \sum_{i=0}^{K} \alpha(K_t, Q_i) \cdot V_{t-i}$$

  with key K_t, query Q_i and intermediate features V. This is a weighting of these features driven by the keys and the queries.
- Here, keys and queries could also be outcomes from some LSTMs, as suggested in the literature.
- Very interesting is that attention mechanisms can learn some **regime-dependent** temporal dynamics which are often important in financial applications.

Hybrid models

- Hybrid models combine existing statistical models with deep-learning techniques
- in this sense, this approach should always outperform the reference method
- We will soon visit some filtering-inspired networks, for which we start by learning what filtering actually is.

 Lim, Bryan and Stefan Zohren (2021). „Time-series forecasting with deep learning: a survey“. In: **Philosophical Transactions of the Royal Society A** 379.2194, p. 20200209.