Multimodal Meta-Learning for Time Series Forecasting

Sebastian Pineda Arango

July 2020

1 Problem formulation

In the proposed thesis, we want to achieve fast adaptation for a regression problem in time series with short-history. This is to be done by finding adequate initialization of the models and by using many time-series (i.e. cross-learning). Consider a model $f_{\tilde{\theta}}$ with parameters $\tilde{\theta}$ that performs regression taken as an input a multivariate time series with C channels and with L observed samples in every channel $f_{\tilde{\theta}} : \mathrm{R}^{L \times C} \rightarrow \mathrm{R}$. We want to find a set of initial parameters $\tilde{\theta}$ using a set of time series $\mathcal{S}^{Tr} = \{\mathbf{s}_i^{Tr} | \mathbf{s}_i^{Tr} \in \mathrm{R}^{L^{Tr} \times C}, i = 1, ..., N^{Tr}\}$, so that it learns fast (few iterations) how to forecast on a new set of time series $\mathcal{S}^{Te} = \{\mathbf{s}_i^{Te} | \mathbf{s}_i^{Te} \in \mathrm{R}^{L^{Te} \times C}, i = 1, ..., N^{Te}, L^{Te} << L^{Tr}\}$.

2 Background

In order to solve a regression problem using a multivariate time series $\mathbf{s}_i \in \mathcal{S}^{Tr}$, we split $\mathbf{s}_i$ in different time windows, using a sliding window. Afterwards, we create a two-element tuple for every window, where the first element is the current window (backcast) and the second element is the target. The set of all the tuples is identified as $W_i = \{(\mathbf{x}_j, \mathbf{y}_j), j = 1, ..., N_i \, \mathbf{x}_j \in \mathrm{R}^L, \mathbf{y}_j \in \mathrm{R}^H\}$. Using this set, we want to learn how predict $\mathbf{y}_i$ given $\mathbf{x}_i$ using $f_{\tilde{\theta}}$. The operator that transforms an original time series $\mathbf{s}_i$ into the set of tuples will be identified as $\mathcal{W}(\cdot)$, such that $W_i = \mathcal{W}(\mathbf{s}_i)$.

Considering the baseline problem of solving the univariate point forecast problem with the model $f_{\tilde{\theta}}$ given a unique time series $\mathbf{s}_i \in \mathcal{S}^{Tr}$, an optimization problem for finding the optimal parameters is formulated as follows:

$$\tilde{\theta}^* = \min_{\tilde{\theta}} \sum_{(\mathbf{x}_j, \mathbf{y}_j) \in W_i} ||\mathbf{y}_j - f_{\tilde{\theta}}(\mathbf{x}_j)||_2^2 \tag{1}$$

Siilarly, we can extend the formulation to a cross-learning, where we optimize for several time-series:

$$\tilde{\theta}^* = \min_{\tilde{\theta}} \sum_{\mathbf{s}_i \sim \mathcal{S}^{Tr}} \mathcal{L}_{\mathbf{s}_i}(f_{\tilde{\theta}}) = \min_{\tilde{\theta}} \sum_{\mathbf{s}_i \sim \mathcal{S}^{Tr}} \sum_{(\mathbf{x}_j, \mathbf{y}_j) \in \mathcal{W}(\mathbf{s}_i)} ||\mathbf{y}_j - f_{\tilde{\theta}}(\mathbf{x}_j)||_2^2 \qquad (2)$$

3 Proposed solution

Meta learning has the goal to train a model so that it can quickly adapt to a new task using few data points. In his master thesis, we propose meta-learning as a way of finding parameters initialization to models so that the learning of time series forecasting problem is faster in time series with short history, as short time series would produce few tuples from the sliding windows (after the $\mathcal{W}$ operator).

In order to apply MAML, the meta-learning framework formulated by (Finn et al., 2017), we propose to treat every time series as a task. Therefore, based on the same paper, we recast the problem to our described problem as:

$$\min_{\tilde{\theta}} \sum_{\mathbf{s}_i \sim \mathcal{S}^{Tr}} \mathcal{L}_{s_i}\left(f_{\tilde{\theta} - \alpha \nabla_{\tilde{\theta}} \mathcal{L}_{\mathbf{s}_i}(f_{\tilde{\theta}})}\right) = \min_{\tilde{\theta}} \sum_{\mathbf{s}_i \sim \mathcal{S}^{Tr}} \sum_{(\mathbf{x}_j, \mathbf{y}_j) \in \mathcal{W}(\mathbf{s}_i)} ||\mathbf{y}_j - f_{\tilde{\theta} - \alpha \nabla_{\tilde{\theta}} \mathcal{L}_{\mathbf{s}_i}(f_{\tilde{\theta}})}(\mathbf{x}_j)||_2^2$$

On top of this and expired by the (MMAML) Multimodal Meta-learning (Risto et al., 2019), we propose to encode the time series $\mathbf{s}_i$ to generate additional parameters τ that modify the main model $f_{\tilde{\theta}}$. If we denote the network that generates τ as h_ϕ such that $\tau_i = h_\phi(\mathbf{s}_i)$, then the final proposed optimization objective looks as follows:

$$\min_{\tilde{\theta}, \phi} \sum_{\mathbf{s}_i \sim \mathcal{S}^{Tr}} \mathcal{L}_{s_i}\left(f_{\tilde{\theta} - \alpha \nabla_{\tilde{\theta}} \mathcal{L}_{\mathbf{s}_i}\left(f_{\tilde{\theta}, \tau_i}\right), \tau_i}\right)$$

The conditining of the main model $f_{\tilde{\theta}}$ using generated parameters τ is denoted simply as $f_{\tilde{\theta}, \tau}$ and uses FiLM layers (Perez et al., 2018) in the same way as MMAML. In order to fit to the literature terminology, we refer to $f_{\tilde{\theta}, \tau}$ as the task network and to h_ϕ as the modulation network. The figure 1 depicts the proposed architecture.

Why meta-learning? Meta-learning allows fast adaptation to new tasks (in this context, new time series) with few data (short history).

Why multi-modal? Time series may present very different distributions in some cases, therefore, conditioning the parameters to the whole time series ($\mathbf{s}_i$ may allow to make better predictions $\hat{\mathbf{y}}_j$ given $\mathbf{x}_j$. Moreover, since $\mathbf{x}_j$ is derived from a smaller time window out of $\mathbf{s}_i$, it has short term information, whereas encoding the whole time series will enable global information. Thus, it yields a hierarchical learning structure.

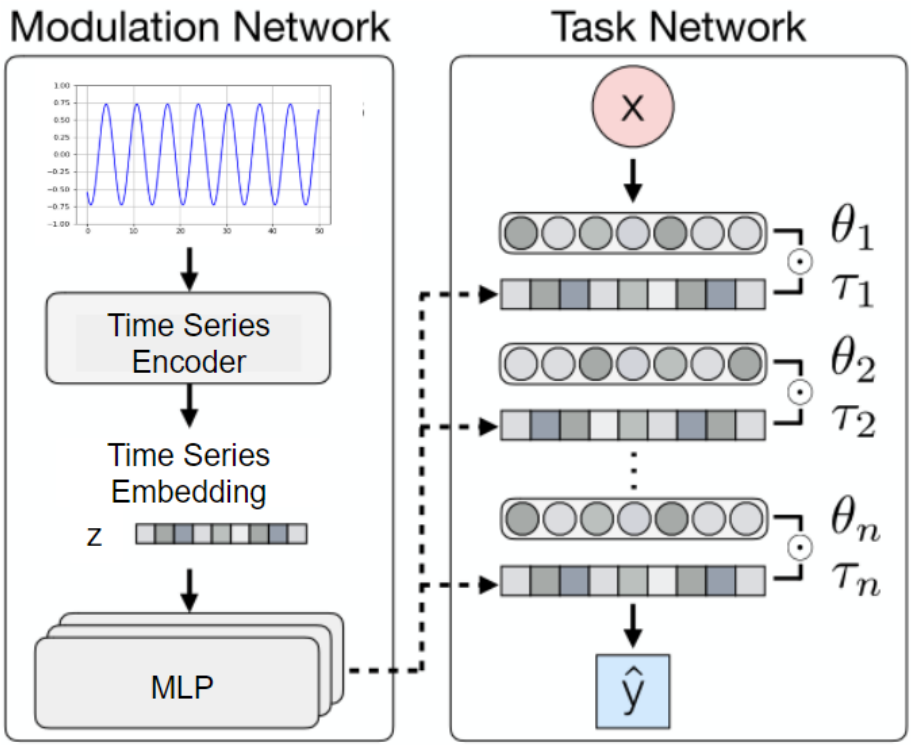

Figure 1: Architecture

References

Vuorio, Risto Sun, Shao-Hua Hu, Hexiang Lim, Joseph. (2019). *Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation.*

Finn, C., Abbeel, P., Levine, S. (2017). *Model-agnostic meta-learning for fast adaptation of deep networks.* arXiv preprint arXiv:1703.03400.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A. (2018, April). *Film: Visual reasoning with a general conditioning layer.* In Thirty-Second AAAI Conference on Artificial Intelligence.