ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVI Cycle

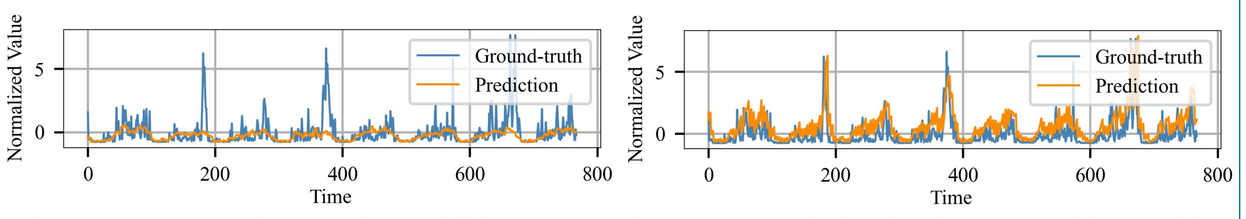
MoQ: A Risk-Adaptive Mixture of Experts for Time Series Forecasting

Shuyang Li

Supervisor: Prof. Enrico Magli

Research context and motivation

- Time series forecasting plays a key role for decision-making in many different domains. For industrial applications, accurate predictions of the upcoming peaks of a time series bring benefits for system management and resource allocation. However, the prediction of potential peaks is extremely difficult due to the fact that many peaks appear suddenly for no apparent reason. Based on this reason, how to improve the peak predictions is seen as a challenging task in time series forecasting field.



- In many cases, time series exhibit a generally stationary behavior interrupted by sudden strong peaks. Many time series forecasting methods are very good at predicting the stationary parts, but provide inaccurate results in the prediction of peaks; on the other hand, these latter often bear a more important practical meaning, making their prediction essential. Generally, a neural network is trained to minimize an average loss on the prediction error; since peaks occur rarely, their prediction tends to be neglected because their effect on the loss is quite small on average, whereas the loss is dominated by errors in the stationary parts of the time series. This leads to models typically making conservative predictions most of the time. In this case, there is always a trade-off between being conservative (good overall performance) and being aggressive (good peak prediction) .

Addressed research questions/problems

- A novel model called MoQ is proposed to tackle the peak prediction problem, which is quite challenging in mobile traffic forecasting fields. A better prediction of mobile traffic would lead to more efficient resource allocation and configuration optimization in mobile networks.
- For the training of Mixture-of-Experts (MoE) model, there are two major difficulties: first it is hard to guarantee that each expert learns different things from the same training data; if all experts learn similar patterns, the benefit of using this framework is minor. Another issue is about expert assignment, whereby the manager may tend to consistently select a specific expert during training which results in an unbalanced expert usage. To address these problems, a two-stage training process is designed carefully.
- The interpretability of neural network has always been a limiting factor for use cases requiring explanations of the results. To improve the interpretability of the forecasting model, we establish a novel experts cooperation mechanism which can be visualized and analyzed to study the behavior of the model, allowing the user to customize the model behavior in a desired way.

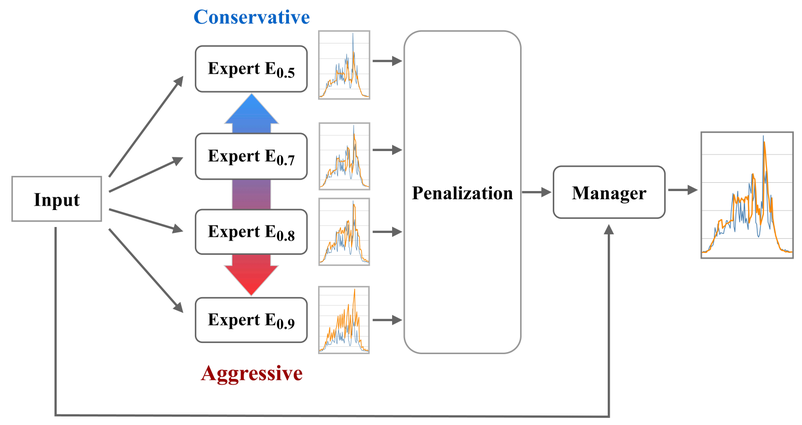
Novel contributions

- A novel model called MoQ is proposed, which supports various forecasting styles, and features a flexible blending of conservative and aggressive predictions based on recent observations.
- A two-stage training process is designed to address the difficulty of training MoE model. Indeed, in MoQ each expert must behave in a specific way. To this end, we first pre-train experts with different objective functions to promote their diversity. In the second stage, penalization is applied during training to prevent the problem of imbalanced assignment of experts and encourage the cooperation among experts.
- Experiments are carried out on real-world datasets, and the results prove that MoQ improves peak prediction significantly. Comparing with baselines, this model is more sensitive to the occurrence of peaks. Visualizing the assigned weights of experts, we observe interpretable cooperation between them, which explains why the model is very effective at adapting to fast changing time series.

Submitted and published works

- [Submitted] Shuyang Li , Gianluca Francini and Enrico Magli, "*Temporal Dynamics Clustering for Analyzing Cell Behavior in Mobile Network*", Computer Networks, Elsevier.
- [Submitted] Shuyang Li , Enrico Magli and Gianluca Francini, "To Be Conservative or To Be Aggressive? A Risk-Adaptive Mixture of Experts for Time Series Forecasting", AAAI Conference 2023.

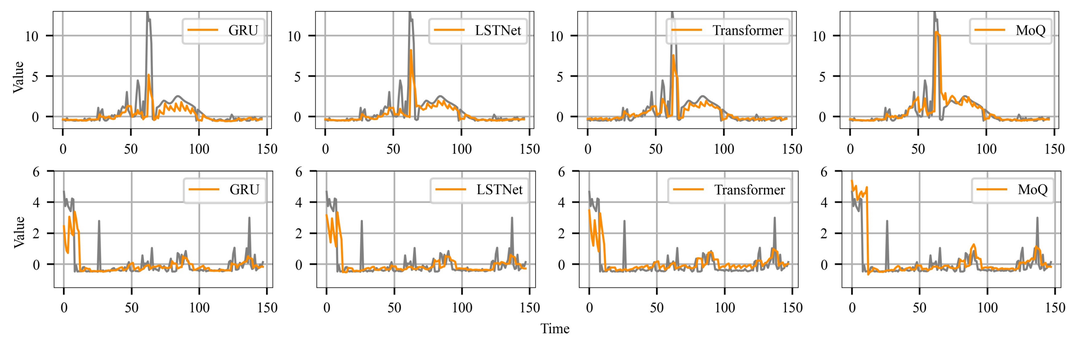
Adopted methodologies



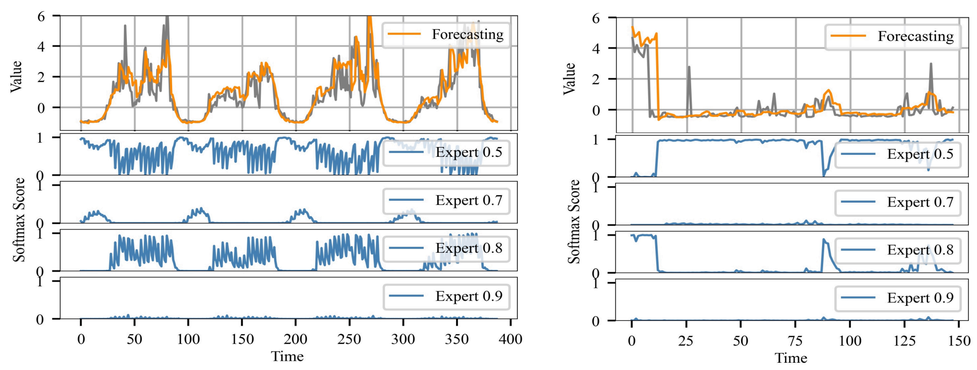
- The idea behind MoQ is to fuse various prediction styles, as illustrated in the top figure. In MoQ, the input is first fed into the experts pool. By training experts with different objective functions, these expert will have diverse forecasting styles: some of them are going to make aggressive predictions while others will be more conservative. How to use these predictions to maximum effect is the job of manager, who observes the recent temporal behavior of the input features and fuses these predictions based on softmax score; in this way, the model automatically learns when to be conservative and when to be aggressive.

Mobile Traffic Dataset

Metrics	MLP	LSTM	GRU	LSTNet	TCN	MQ-RNN	Transformer	MoQ	$MoQ_2^\dagger$	$MoQ_4^\dagger$
MAE	0.304	0.295	0.292	**0.288**	0.395	0.299	0.325	0.312	0.302	0.319
MSE	0.289	0.296	0.290	**0.283**	0.448	0.305	0.326	0.337	0.295	0.322
Accuracy	68.3%	68.1%	67.9%	66.7%	59.5%	64.7%	67.1%	75.9%	74.3%	**76.4%**
Sensitivity	37.7%	37.1%	36.6%	34.1%	19.5%	30.1%	35.0%	54.5%	50.6%	**55.7%**



- Forecasting is performed on a real-world mobile traffic dataset which covers 100 cells, and the performance of the models is reported in the table. Comparing with the selected baselines, the proposed MoQ models obtain much higher sensitivity of detecting peaks. Among these MoQ models, $MoQ_4^\dagger$ has the highest sensitivity whose value is 18% higher than the highest sensitivity of baselines (37.7%), which means this model is more capable of predicting potential peaks in upcoming steps; it also has the highest classification accuracy, that means the proposed model is able to perform forecasting based the recent trend instead of overestimating the targets all the time. The price to be paid for very good sensitivity is that its MAE and MSE are slightly higher than those of LSTNet, though still to close to those achieved by the best methods.



Future work

- For the future work, it is interesting to see if this model can be used to detect the poential mobile netowork congestion in near future or to be integraed with the state-of-the-art reforcement learning model to improve the perofrmance of configuration optimization.

List of attended classes

- 01VPORW – Statistical methods with application to climate variability and change assessments (10/06/2022, 5 CFU)
- 01SCTIU – Text mining and analytics (30/09/2021, 3 CFU)
- 01SCVIU – Data analytics for science and society (30/09/2021, 3 CFU)
- 01UJBRV – Adversarial training of neural networks (03/06/2021, 3 CFU)
- 01UMNRV – Advanced deep Learning (15/06/2021, 6 CFU)

Politecnico di Torino

PhD program in
Electrical, Electronics and Communications Engineering