Towards an Optimal Management of the 5G Cloud-RAN through a Spatio-Temporal Prediction of Users' Dynamics

Arcangela Rago[1,2], Pasquale Ventrella[1], Giuseppe Piro[1,2], Gennaro Boggia[1,2], Paolo Dini[3]
[1]*Dept. of Electrical and Information Engineering, Politecnico di Bari,* Italy
Email: {arcangela.rago, pasquale.ventrella, giuseppe.piro, gennaro.boggia}@poliba.it
[2]*CNIT, Consorzio Nazionale Interuniversitario per le Telecomunicazioni,* Italy
[3]*CTTC, Centre Tecnològic de Telecomunicacions de Catalunya,* Spain
Email: {paolo.dini}@cttc.es

***Abstract*—In the emerging 5G architecture, the Cloud-Radio Access Network (Cloud-RAN) offers the possibility to dynamically configure virtual resources and network functionalities very close to end-users, while jointly considering bandwidth, computing, latency, and memory capabilities requested by heterogeneous applications, the channel quality experienced by end-users, mobility, and any kind of system constraints. By capitalizing on recent scientific results and standardization activities on 5G, this short paper presents a preliminary design of an ETSI-NFV compliant architecture willing to support the implementation of advanced protocols, algorithms, and methodologies for the optimal management of the 5G Cloud-RAN. Its components and functionalities have been sketched by harmoniously integrating Software-Defined Networking (SDN) facilities, Multi-access Edge Computing (MEC), and deep learning. Herein, spatio-temporal users' dynamics are collected by SDN controllers and predicted by a high-level orchestrator through a Convolutional Long Short-Term Memory scheme. Then, the outcomes of the prediction process are adopted to dynamically configure the Cloud-RAN (i.e., by using any kind of customizable algorithm). Some of the capabilities of the proposed approach are preliminarily evaluated by considering the autonomous driving use case and real mobility traces. Moreover, the paper concludes by reporting an overview of future directions of this research activity.**

***Index Terms*—5G Cloud-RAN, Users' dynamics, ConvLSTM**

I. Introduction

With the explosive growth of communication traffic and the arrival of the fifth generation (5G) of mobile broadband systems, traffic and mobility prediction are needed for an effective planning and usage of network resources [1], [2]. In this context, deep learning could be properly tailored to anticipate traffic behaviors and optimize the deployment of virtual resources and functionalities very close to end-users (i.e., at the edge of the network), while offering concrete answers to the deployment of flexible and advanced applications asking for bandwidth, computing, latency, and memory capabilities never seen before [3], [4].

The current scientific literature generally investigates traffic forecasting and mobility prediction separately. The prediction of the mobile traffic load has been achieved through Convolutional Neural Networks (CNNs) [5], Long Short-Term Memorys (LSTMs) [6]–[8], or a combination of them [9], [10]. Mobility prediction is achieved through Markov Chains [11], Markov Decision Processes [12], Hidden Markov Models [13], Bayesian Networks [14], Neural Networks [3], [15], [16], or a combination of Neural Networks and Bayesian Networks [17].

Differently from the current state of the art, this short paper jointly addresses the two aforementioned problems and conceives a network architecture willing to optimally manage the 5G Cloud-Radio Access Network (Cloud-RAN) through deep learning [18]. The high variability and heterogeneity of components and functionalities that compose the conceived framework inevitably make the design of a suitable deep learning algorithm a very challenging task to accomplish. Therefore, an original methodology leverages the integration of Software-Defined Networking (SDN) facilities, Multi-access Edge Computing (MEC), and deep learning is sketched in support of a preliminary resource planning through the prediction of spatio-temporal users' dynamics. It is important to note that at the time of writing, and to the best of our knowledge, a first attempt in this direction is presented in [19]. Here, a multivariate LSTM is developed for predicting the workload in MEC entities, by considering the impact of user mobility. This short paper significantly advances the current state of the art, including [19], because: i) it frames the overall proposal within the standardized ETSI-NFV architecture, ii) it proposes a new methodology for the spatio-temporal prediction of users' dynamics (which differs from the one adopted in [19]), and iii) it provides a very preliminary discussion on the usage of prediction outcomes in a realistic use case.

The remainder of the short paper is as follows. Section II illustrates the proposed architecture and provides some technical details on the adopted deep learning approach. Section III presents the preliminary investigation, including the processed data and the early results. Finally, Section IV concludes the paper and draws future research activities.

II. The proposed Architecture

The network architecture presented herein wants to natively support a wide range of services, including autonomous driving, augmented reality, virtual reality, and drones (just to name a few), that have massive bandwidth and latency constraints. In line with 5G specifications, gNBs provide wireless connectivity to mobile users through heterogeneous technical components at the radio interface (this concept is illustrated in Fig. 1 by means of beams with different colors) [20]. A number of MEC servers are connected to gNBs and expose computing resources to mobile users,

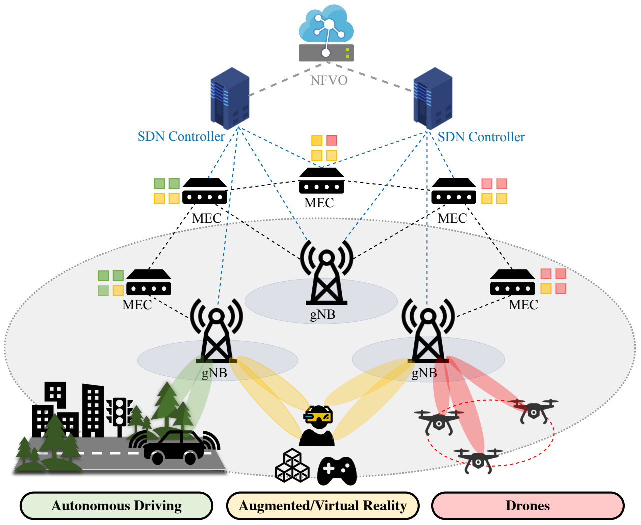


Fig. 1. The proposed architecture.

depending on the service they use [21]. Also in this case, the colored blocks close to MEC servers in Fig. 1 highlight the heterogeneity of resources allocated for different applications. All the resources available in the Cloud-RAN are monitored, configured, and orchestrated according to the ETSI-NFV architecture [18]. Specifically, gNBs and MEC servers are connected to SDN controllers. They locally control users' mobility and monitor network resources requested by the mobile users. The NFV Orchestrator (NFVO) optimally orchestrates network services and resources on the Cloud-RAN, based on the prediction of spatio-temporal users' dynamics, while satisfying heterogeneous traffic demands [21]. Note that radio and MEC resources can be dynamically allocated to a group of services, according to the network slice paradigm [22].

The main functionalities covered by the proposed architecture are introduced below. Only a high-level description is presented and their complete design is delayed for future research activities.

A. *Monitoring of users' mobility and resource usage*

Each SDN controller implements monitoring functionalities and retrieves spatio-temporal users' dynamics, including mobility patterns and bandwidth utilization. The interaction between SDN controller and the other entities of the network is implemented through conventional communication control protocols (i.e., OpenFlow, RestConf, etc.) [23]. However, the structure of provided data (the YANG model, for instance) and the periodicity of that interaction remain an open issue and must be properly defined.

B. *Recognition of user mobility patterns*

The key methodology envisaged in this contribution assumes to predict the spatio-temporal users' dynamics through deep learning. In fact, spatio-temporal users' dynamics captured by SDN controllers are collected by the NFVO, which can consequently perform mobility prediction. Specifically, the *Convolutional LSTM (ConvLSTM)* architecture, which has been initially introduced for precipitation nowcasting [24] and recently investigated also for traffic forecasting [25], is adopted for this purpose. The ConvLSTM is a neural network based on LSTM [26], with the convolution operator as input, forget, and output gates instead of the element-wise or Hadamard product [24]. Therefore, it can extract temporal and spatial correlations of data through LSTM memory cells and the convolutional operation, respectively [10], [27]. Going more into detail, this work conceives a learning architecture embracing two 2-dimensional ConvLSTM layers, after each one a *Batch Normalization* layer is used to accelerate deep network training [28]. At the end, the prediction is performed through a fully-connected layer with the Rectified Linear Unit (ReLU) activation function [10]. The predictor is configured in order to minimize the Mean Square Error (MSE) loss function. The distribution of users among cells and the resources they use at both radio interface and Cloud-RAN (also on the network slice bases) are observed for a time interval T. Then, the ConvLSTM is used to predict these details in the future time instants.

The dimension of cells, the observation slot, and the duration of T are relevant parameters for future research activities. Moreover, the robustness of the prediction algorithm to deal with uncertainty and measurement errors has to be considered in the design and evaluated. Another important aspect is the algorithm complexity together with the availability of training data. It is recommended not to send a huge amount of data through wireless links and avoid congestions. In this context, distributed learning solutions must be studied to share knowledge among the different MEC servers.

C. *Towards an optimal resource management*

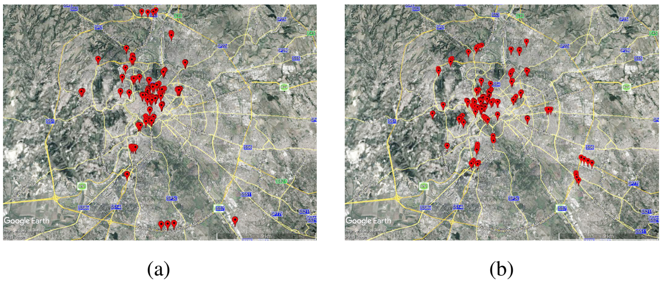
The outcomes of the prediction process are adopted to dynamically configure the 5G Cloud-RAN. In this case, user mobility patterns may be used to aid optimization algorithms to allocate radio resources among network slices, initiate or configure MEC resources based on users' demands [21]. For example, NFVO may forecast next user locations and take full advantage of good future conditions (such as getting closer to a gNB or entering a less loaded MEC server) or mitigate the impact of negative events (e.g., entering a tunnel). A careful study on the impact of prediction error on the optimization problem needs further investigations. It might be potentially more harmful to use a wrong prediction than not using prediction at all. A good accuracy can usually be obtained for short prediction horizons, which, however, should be of a correct length to make the optimization algorithms benefit from it. Therefore, a good balance between prediction horizon and accuracy must be found.

III. Preliminary Investigation

The preliminary results discussed in this position paper refer to the prediction functionality presented in Section II-B. The autonomous driving use case is considered as an example and the distribution of mobile users in the spatio-temporal domain is given by realistic mobility traces.

A. *Dataset*

This short paper considers the dataset presented in [29], which reports the movements of 316 taxi cabs in the center of Rome, from 1 February 2014 to 2 March 2014, with a granularity of about 15s. Fig. 2 shows an example of the taxi distribution at 1:00 pm and 1:59 pm. The considered geographical area of around $110km^2$ is bounded by the coordinates pairs (41.793363, 12.372258) (41.991896, 12.616472). It has been divided using 11×10 square cells, so that each grid cell covers a square area of $1km \times 1km$. Therefore, the training dataset has been conveniently pre-processed to be managed by the adopted deep learning architecture. The traces are used to generate a temporal sequence, with a time granularity of 1s, of matrices, whose elements represent the number of taxi in one of the 110 square cells.


(a)

(b)

Fig. 2. Example of taxi distribution at (a) 1:00 pm and (b) 1:59 pm in Rome.

B. Evaluation Setup

The conceived architecture has been implemented in Keras, a high-level neural networks API written in Python, running on top of TensorFlow [30]. The observation window T of the spatio-temporal dynamics is set to 20s. The Adam optimization, with a learning rate equal to 0.001, is used to iteratively update the network weights. The other training hyperparameters, that have been chosen for the scheme implementation, are set as follows: *number of filters* = 200, *kernel size* = 3×3, *number of epochs* = 30, and *batch size* = 64. To preliminary evaluate the mobility prediction, we select the daily time slot from 1:00 pm to 1:59 pm as an example of hour with peak taxi activity.

C. Mobility prediction

To evaluate the prediction performance of the conceived approach, we select two significant cells (i.e. cell ID 45 and 55) as examples to plot the observed and the predicted trends over time of spatio-temporal users' dynamics. Fig. 3 shows the observed and the predicted trends over time of spatio-temporal users' dynamics for two significant cells. In particular, the blue solid line represents the ground truth of the number of users, while the red dashed line describes the predicted number of users, that are rounded up to the nearest integer. It can be noted that the two trends are almost overlapped. Fig. 4 reports the Mean Absolute Error (MAE) values for the different cell IDs. Generally, MAE is lower than 0.6; only a few cells present peaks equal to 0.8.

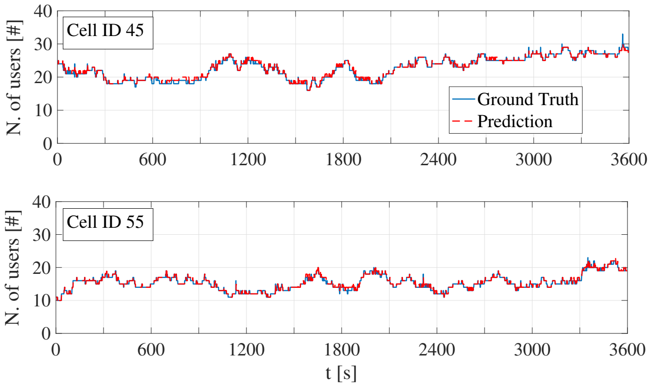


Fig. 3. Prediction of the number of users for example two cells.

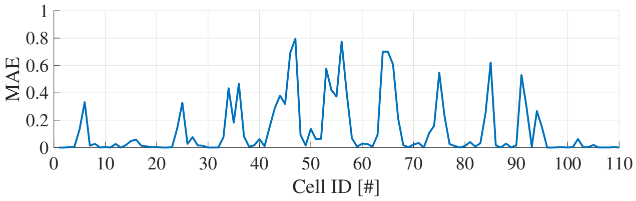


Fig. 4. MAE for each cell

D. Resource Planning

Services for the autonomous vehicles require 16 GB Synchronous Dynamic Random Access Memory (SDRAM) and 100 Mbps as bandwidth [18], [31]. Knowing the spatio-temporal users' dynamics and the minimum requirements of autonomous vehicles, we can preliminary estimate the overall radio and computing resources to be allocated in each cell, according to the following relation: $\hat{R} = \hat{N_j} \cdot r$, where $\hat{N_j}$ is the predicted number of users in the j-th cell and r is the resource requirement in the Cloud-RAN. Fig. 5 shows the actual and the predicted resources in the example two cells, i.e. cell ID 45 and 55. The predicted resources' trend follows the number of users in the cell due to the basic multiplicative estimation proposed in this short paper. As previously anticipated, the actual and the predicted trends are almost overlapped. Therefore, the conceived architecture has good prediction performance of spatio-temporal users' dynamics and resource requests.

IV. Conclusions

This work has preliminarily presented the design of an ETSI-NFV compliant architecture that can optimally manage the 5G Cloud-RAN. Its components and functionalities have been sketched, with a focus on mobility prediction. In fact, spatio-temporal users' dynamics have been predicted through a Convolutional Long Short-Term Memory scheme by considering one-hour mobility traces. Then, the outcomes of the prediction process have been used to quantify the resources to allocate in the Cloud-RAN for the autonomous driving use case. Further research activities will investigate the interaction between Software-Defined Networking controllers and the other entities of the network. Moreover, we will analyze the prediction approach with different configuration parameters and distributed learning solutions. Then, mobility prediction, with a trade-off between prediction horizon and accuracy, could aid optimization algorithms to dynamically configure the 5G Cloud-RAN.

Acknowledgment

This work was supported by the PRIN project no. 2017NS9FEY entitled ”Realtime Control of 5G Wireless Networks: Taming the Complexity of Future Transmission and Computation Challenges” funded by the Italian MIUR, by the Apulia Region (Italy) Research project INTENTO (36A49H6), by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675891 (SCAVENGE), and by Spanish MINECO grant TEC2017-88373-R (5G-REFINE). It has been also partially supported by

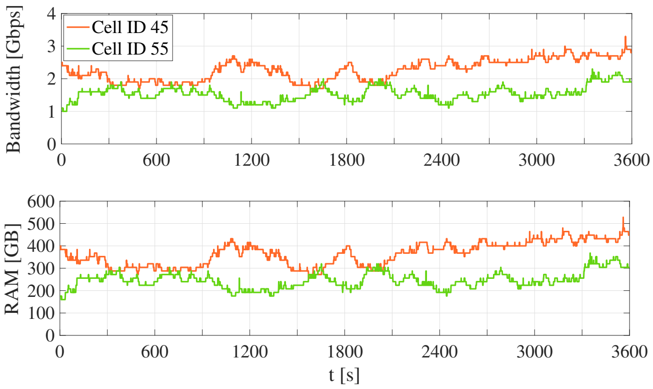


Fig. 5. Estimation of (a) radio resources and (b) MEC resources for cell ID 45 and 55.

the Italian MIUR PON projects Pico&Pro (ARS01_01061), AGREED (ARS01_00254), FURTHER (ARS01_01283), and RAFAEL (ARS01_00305).

References

[1] A. Azari, P. Papapetrou, S. Denic, and G. Peters, "User Traffic Prediction for Proactive Resource Management: Learning-Powered Approaches," *arXiv preprint arXiv:1906.00951*, 2019.

[2] B. Ma, W. Guo, and J. Zhang, "A Survey of Online Data-Driven Proactive 5G Network Optimisation Using Machine Learning," *IEEE Access*, vol. 8, pp. 35 606–35 637, 2020.

[3] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.

[4] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[5] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656–1659, 2018.

[6] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li *et al.*, "Deep mobile traffic forecast and complementary base station clustering for C-RAN optimization," *Journal of Network and Computer Applications*, vol. 121, pp. 59–69, 2018.

[7] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, "DeepTP: An End-to-End Neural Network for Mobile Cellular Traffic Prediction," *IEEE Network*, vol. 32, no. 6, pp. 108–115, 2018.

[8] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep Learning with Long Short-Term Memory for Time Series Prediction," *IEEE Communications Magazine*, 2019.

[9] C. Zhang, M. Fiore, and P. Patras, "Multi-Service Mobile Traffic Forecasting via Convolutional Long Short-Term Memories," in *2019 IEEE International Symposium on Measurements Networking (M&N)*, July 2019, pp. 1–6.

[10] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, Thirdquarter 2019.

[11] S. H. Ariffin, N. Abd, N. E. Ghazali *et al.*, "Mobility prediction via Markov model in LTE femtocell," *International Journal of Computer Applications*, vol. 65, no. 18, 2013.

[12] J. Plachy, Z. Becvar, and E. C. Strinati, "Dynamic Resource Allocation Exploiting Mobility Prediction in Mobile Edge Computing," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2016, pp. 1–6.

[13] A. Magnano, X. Fei, and A. Boukerche, "Movement prediction in vehicular networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–6.

[14] Y. Zhang, J. Hu, J. Dong, Y. Yuan, J. Zhou, and J. Shi, "Location prediction model based on Bayesian network theory," in *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*. IEEE, 2009, pp. 1–6.

[15] C. Wang, L. Ma, R. Li, T. S. Durrani, and H. Zhang, "Exploring trajectory prediction through machine learning methods," *IEEE Access*, vol. 7, pp. 101 441–101 452, 2019.

[16] H. Zhang and L. Dai, "Mobility prediction: A survey on state-of-the-art schemes and future applications," *IEEE Access*, vol. 7, pp. 802–822, 2018.

[17] O. Narmanlioglu, E. Zeydan, M. Kandemir, and T. Kranda, "Prediction of Active UE Number with Bayesian Neural Networks for Self-Organizing LTE Networks," in *2017 8th International Conference on the Network of the Future (NOF)*. IEEE, 2017, pp. 73–78.

[18] ETSI, "Cloud RAN and MEC: A Perfect Pairing ," *White Paper*, no. 23, February 2018.

[19] C. Nguyen, C. Klein, and E. Elmroth, "Multivariate LSTM-based Location-aware Workload Prediction for Edge Data Centers," in *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2019, pp. 341–350.

[20] 5G PPP Architecture Working Group, "View on 5G Architecture," *White Paper*, no. 3, June 2019.

[21] ETSI, "Deployment of Mobile Edge Computing in an NVF environment," *ETSI GR MEC*, no. 17, February 2018.

[22] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, 2020.

[23] W. Queiroz, M. A. Capretz, and M. Dantas, "An Approach for SDN Traffic Monitoring Based on Big Data Techniques," *Journal of Network and Computer Applications*, vol. 131, pp. 28–39, 2019.

[24] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A Machine Learning Approach for Precipitation Nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[25] Y. Ma, Z. Zhang, and A. Ihler, "Multi-Lane Short-Term Traffic Forecasting with Convolutional LSTM Network," *IEEE Access*, pp. 1–1, 2020.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[28] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015.

[29] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD dataset roma/taxi (v. 2014-07-17)," Downloaded from https://crawdad.org/roma/taxi/20140717/taxicabs, Jul. 2014.

[30] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[31] M. Jung, S. A. McKee, C. Sudarshan, C. Dropmann, C. Weis, and N. Wehn, "Driving into the Memory Wall: the Role of Memory for Advanced Driver Assistance Systems and Autonomous Driving," in *Proceeding of the International Symposium on Memory Systems*, 2018, pp. 377–386.