Designing Deep Learning Model for Accurate vNF Service Chain Pre-deployment in Inter-DC EONs

Baojia Li, Wei Lu, Siqi Liu, Zuqing Zhu†
School of Information Science and Technology, University of Science and Technology of China, Hefei, China
†Email: {zqzhu}@ieee.org

***Abstract*—This work considers a provisioning framework for inter-datacenter elastic optical networks, which incorporates prediction and pre-deployment of virtual network function service chains (vNF-SCs), and designs a new deep learning (DL) model to realize accurate prediction. Simulation results confirm that the new DL model provides higher prediction accuracy than the existing approach, and with it, a service provider achieves better tradeoff between resource utilization and blocking probability.**

***Index Terms*—Network function virtualization, Service chaining, Deep learning, Elastic optical networks (EONs).**

I. Introduction

In today's datacenters (DCs), service providers (SPs) leverage IT resource virtualization to realize network function virtualization (NFV) [1, 2] and deploy various virtual network functions (vNFs) timely over general-purpose servers, switches and storages. Meanwhile, the creation of new services can be further expedited by steering traffic through a series of vNFs, *i.e.*, realizing a vNF service chain (vNF-SC) [3]. On the other hand, the rapid growth of data-/bandwidth-intensive and real-time services in the Internet makes the capacity and flexibility of the underlying inter-DC networks essential for achieving high-performance vNF-SC provisioning. Therefore, people try to explore the agile bandwidth allocation provided by the flexible-grid elastic optical networks (EONs) [4] and build inter-DC networks over EONs (IDC-EONs) [5].

Note that, in order to realize on-demand and cost-effective vNF-SC provisioning in an IDC-EON, an SP needs to properly address the challenge due to setup latency [6]. Specifically, the joint optimization of spectrum and IT resource allocations in the IDC-EON would be complex and time-consuming [5], and the relatively long configuration latencies from lightpath establishment [7] and vNF deployment could make real-time and on-demand vNF-SC provisioning infeasible. Hence, in our previous study in [6], we designed a provisioning framework with vNF-SC pre-deployment to resolve the challenge.

In the framework, network operations for vNF-SC provisioning are performed periodically in fixed time slots (TS'), where each TS includes two phases, *i.e.*, the pre-deployment and provisioning phases. Each TS starts with the pre-deployment phase, in which the SP first uses a deep learning (DL) model to predict future vNF-SC requests in the TS and then performs lightpath establishment and vNF deployment accordingly. Next, in the provisioning phase, the SP serves actual arrival requests almost immediately by steering their traffic through the required vNFs in sequence. This framework successfully addresses the challenge due to setup latency, because the latencies from the joint optimization of resource allocations, lightpath establishment, and vNF deployment are removed from the setup latency of a vNF-SC, while with software-defined networking (SDN), traffic steering can be accomplished within hundreds of milliseconds [8].

Although we designed a DL model based on the long/short-term memory based neural network (LSTM-NN) [9] to forecast the high-dimensional data of future vNF-SC requests[1], the DL model has not been fully optimized to achieve high prediction accuracy. This motivates us to redesign the DL model for realizing more accurate vNF-SC pre-deployment in this work. Specifically, we propose two modifications to the DL model, which are 1) a novel scheme that encodes the vNF-SC of each request as a feature matrix to better represent the similarity and difference among vNF-SCs, and 2) the addition of abstraction and concretion layers in the LSTM-NN to acquire more accurate predictions. Simulation results show that the newly-designed DL model provides higher prediction accuracy than the one in [6], and with it, an SP can achieve lower blocking probability in the provisioning phase.

The rest of the paper is organized as follows. Section II presents the problem formulation that describes the network model and vNF-SC provisioning framework. The design of the new DL model is explained in Section III, and we discuss the numerical simulations for performance evaluation in Section IV. Finally, Section V summarizes the paper.

II. Problem Formulation

We model the IDC-EON as a directed graph $G(V, E)$, where V and E are the sets of DC nodes and fiber links, respectively. A DC node is the abstraction of a DC for vNF deployment and a bandwidth-variable optical cross-connect (BV-OXC) for establishing inter-DC lightpaths. The IT resource capacity of the DC in DC node $v \in V$ is C_v, while each fiber link $e \in E$ can accommodate F frequency slots (FS'). The bandwidth of an FS is assumed to be 12.5 GHz, which can carry 12.5 Gbps traffic throughput [10]. The SP can provision M types of vNFs in the DCs, where an m-th type vNF (*i.e.*, *vNF-m*) consumes c_m units of IT resources and handles a peak traffic throughput of b_m Gbps. We represent a vNF-SC as $\{f_1, f_2, \cdots, f_K\}$, where f_k is the k-th vNF in the vNF-SC and K denotes the

[1]A vNF-SC request is modeled with its source-destination pair, bandwidth requirement, arrival time, hold-on time, and vNF sequence.

number of vNFs in the longest vNF-SC that can be supported in the IDC-EON. Then, we can model a dynamic vNF-SC request as $R^i = \{s^i, d^i, \{f_1, f_2, \cdots, f_K\}^i, b^i, t_a^i, t_h^i\}$, where i is its unique index, s^i and d^i are the source and destination DC nodes, $\{f_1, f_2, \cdots, f_K\}^i$ is the required vNF-SC, b^i is its bandwidth requirement in Gpbs, and t_a^i and t_h^i are its arrival time and hold-on time, respectively [6].

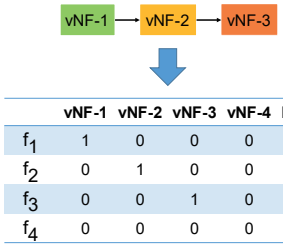
We consider the IDC-EON as a discrete-time system [6], where the SP serves vNF-SC requests in fixed TS', each of which has a duration of ΔT and includes a pre-deployment phase followed by a provisioning phase. In the pre-deployment phase, the SP first predicts the vNF-SC requests that will arrive in the provisioning phase based on historical requests with a DL model. Then, for each predicted request in ascending order of arrival time, the SP deploys the required vNFs on DCs and sets up lightpaths to connect the vNFs in sequence, using the LBA algorithm developed in [3]. Finally, the SP removes idle lightpaths and vNFs to improve resource utilization. Next, the network system enters the provisioning phase, in which the SP tries to steer client traffic through pre-deployed vNFs and lightpaths to formulate vNF-SCs immediately upon receiving a vNF-SC request. Here, a request can be blocked if it cannot be served with the pre-deployed resources. In this framework, the prediction accuracy of the DL model is the key to achieve high performance vNF-SC provisioning, *i.e.*, the best tradeoff between resource utilization and blocking probability.

III. Design of Improved DL Model

The new DL model follows the similar operation principle proposed in [6]. Specifically, it uses a prediction window whose size is w requests, and a future request R^{I+1} is obtained by analyzing the $w+1$ latest historical requests (*i.e.*, $\mathbf{R_h} = \{R^i,\ i \in [I-w, I]\}$). Then, R^{I+1} is treated as the latest historical request and the DL model moves the prediction window forward for one request (*i.e.*, $\mathbf{R_h} = \{R^i,\ i \in [I-w+1, I+1]\}$) to predict R^{I+2}. This procedure is repeated in each pre-deployment phase until all the requests that will arrive in the subsequent provisioning phase have been predicted. We propose the following two modifications to improve the DL model's prediction accuracy.

A. Encoding Scheme of vNF-SCs

To facilitate accurate prediction in the DL model, we have to encode the six dimensional parameters in each request $R^i = \{s^i, d^i, \{f_1, f_2, \cdots, f_K\}^i, b^i, t_a^i, t_h^i\}$ properly. However, we find that the scheme designed for encoding vNF-SCs in [6] would have scalability issues and introduce uncontrollable prediction errors. This is because we assumed that there would be N types of vNF-SCs in the IDC-EON, and encoded $\{f_1, f_2, \cdots, f_K\}^i$ as an N-element vector. In the worst case, each f_k can take any of the M types of vNFs and N can be as large as M^K, which clearly would be neither efficient nor scalable. Moreover, the encoding scheme makes the Euclidean distances between any two different N-element vectors be the same, *i.e.*, the difference and similarity between any two different vNF-SCs are the same. This would lead to uncontrollable prediction errors since the difference and similarity between two different vNF-SCs are different.



	vNF-1	vNF-2	vNF-3	vNF-4	None
f_1	1	0	0	0	0
f_2	0	1	0	0	0
f_3	0	0	1	0	0
f_4	0	0	0	0	1

Fig. 1. Example of vNF-SC encoding.

To overcome the drawbacks mentioned above, we encode a vNF-SC $\{f_1, f_2, \cdots, f_K\}$ as a feature matrix $\mathbf{SC}_{K,M+1}$ in this work, with K rows and $M+1$ columns. Here, element (k, m) equals 1 if the k-th vNF in the vNF-SC is an m-th type vNF, and 0 otherwise. The $(M+1)$-th column corresponds to the unused cases, for correctly modeling the vNF-SCs whose lengths are shorter than K. Fig. 1 gives an example on the vNF-SC encoding, where $K = 4$ and $M = 4$. The feature matrix is for vNF-SC: *vNF-1→vNF-2→vNF-3*. Since the vNF-SC only consists of three vNFs, element $(4, 5)$ equals 1 to indicate that the fourth vNF in the vNF-SC is unused.

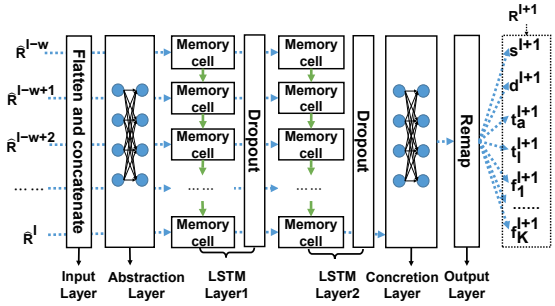


Fig. 2. Design of improved LSTM-NN for vNF-SC request prediction.

B. Improved LSTM-NN

To ensure that the LSTM-NN can adapt to the new encoding scheme and provide higher prediction accuracy, we redesign it as shown in Fig. 2. The input layer takes $w+1$ latest historical encoded requests as the inputs, *e.g.*, $\hat{\mathbf{R}}_\mathbf{h} = \{\hat{R}^i,\ i \in [I-w, I]\}$. As each encoded request contains a feature matrix for its vNF-SC and five variables/vectors for the remaining parameters, the input layer flattens the feature matrix as a long vector and concatenates it with the remaining parameters to form a vector. Then, the vectors of all the historical requests are forwarded to the abstraction layer, where the vectors are analyzed by two fully-connected sub-layers to extract the correlations and features in them. Next, after receiving the outputs from the abstraction layer, LSTM-Layer1 updates the states of its memory cells accordingly and passes the information to be further processed in LSTM-Layer2. The green and blue dashed arrows in Fig. 2 show how the information gets

transferred and the states get updated, respectively. Here, we add a random dropout module after each LSTM layer to avoid over-fitting. After LSTM-Layer2, the concretion layer uses a similar structure as that of the abstraction layer to convert obtained information to predicted request parameters. Finally, the output layer remaps the predicted parameters to a future vNF-SC request, *e.g.*, R^{I+1}.

IV. Performance Evaluation

The simulations use the same scheme as that in [6] to generate the vNF-SC requests based on the traces for real wide-area TCP connections in [11]. We totally generate around $50,000$ dynamic vNF-SC requests and put 80% and 20% of them into the training and testing sets, respectively. The performance evaluation compares the newly-designed improved DL model (Improved-DL) with the DL model (DL) developed in [6].

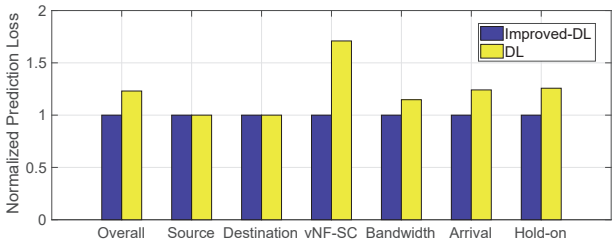


Fig. 3. Comparison of prediction loss on testing set.

We first compare the prediction loss of Improved-DL and DL, and Fig. 3 shows the results on normalized prediction losses verified with the testing set for different parameters. The overall prediction loss from Improved-DL is smaller than that of DL. The reduction is mainly achieved by reducing the prediction loss on vNF-SCs significantly, which verifies the effectiveness of the proposed vNF-SC encoding scheme. Meanwhile, since the new encoding scheme and Improved-DL work together to extract more information from the historical vNF-SC requests, the prediction errors on other parameters also get reduced.

Fig. 4. Results on blocking probability in provisioning phase.

Then, we conduct simulations on dynamic vNF-SC provisioning to further evaluate our proposal. The simulations still use an IDC-EON with the 14-node NSFNET topology [12], and their settings are similar to those in [6]. We apply two DL models in the pre-deployment phases for request prediction, and then plot the results on blocking probability and resource utilization in the provisioning phase in Figs. 4 and 5, respectively. Here, the traffic load in each figure refers to the number of requests arriving in each TS. We observe that Improved-DL can provides lower blocking probability than DL, while their results on resource utilization are almost the same. This confirms that the provisioning algorithm with Improved-DL achieves better tradeoff between resource utilization and blocking probability.

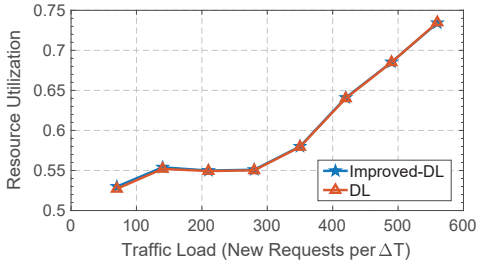


Fig. 5. Results on resource utilization in provisioning phase.

V. Conclusion

We designed a new DL model for realizing more accurate vNF-SC pre-deployment in IDC-EONs. Simulation results suggested that the new DL model can provide higher prediction accuracy than the existing approach, and with it, an SP can achieve better tradeoff between resource utilization and blocking probability in the provisioning phase.

References

[1] "Network functions virtualization (NFV)," Jan. 2012. [Online]. Available: https://portal.etsi.org/portal/server.pt/community/NFV/367

[2] M. Zeng, W. Fang, and Z. Zhu, "Orchestrating tree-type VNF forwarding graphs in inter-DC elastic optical networks," *J. Lightw. Technol.*, vol. 34, pp. 3330–3341, Jul. 2016.

[3] W. Fang *et al.*, "Joint spectrum and IT resource allocation for efficient vNF service chaining in inter-datacenter elastic optical networks," *IEEE Commun. Lett.*, vol. 20, pp. 1539–1542, Aug. 2016.

[4] Z. Zhu, W. Lu, L. Zhang, and N. Ansari, "Dynamic service provisioning in elastic optical networks with hybrid single-/multi-path routing," *J. Lightw. Technol.*, vol. 31, pp. 15–22, Jan. 2013.

[5] W. Fang *et al.*, "Joint defragmentation of optical spectrum and IT resources in elastic optical datacenter interconnections," *J. Opt. Commun. Netw.*, vol. 7, pp. 314–324, Mar. 2015.

[6] B. Li, W. Lu, S. Liu, and Z. Zhu, "Deep-learning-assisted network orchestration for on-demand and cost-effective vNF service chaining in inter-DC elastic optical networks," *J. Opt. Commun. Netw.*, vol. 10, pp. D29–D41, Oct. 2018.

[7] J. Yin *et al.*, "Experimental demonstration of building and operating QoS-aware survivable vSD-EONs with transparent resiliency," *Opt. Express*, vol. 25, pp. 15 468–15 480, 2017.

[8] S. Li *et al.*, "Improving SDN scalability with protocol-oblivious source routing: A system-level study," *IEEE Trans. Netw. Serv. Manag.*, vol. 15, pp. 275–288, Mar. 2018.

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

[10] Z. Zhu *et al.*, "Demonstration of cooperative resource allocation in an OpenFlow-controlled multidomain and multinational SD-EON testbed," *J. Lightw. Technol.*, vol. 33, pp. 1508–1514, Apr. 2015.

[11] V. Paxson, "Empirically derived analytic models of wide-area TCP connections," *IEEE/ACM Trans. Netw.*, vol. 2, pp. 316–336, Aug. 1994.

[12] P. Lu and Z. Zhu, "Data-oriented task scheduling in fixed- and flexible-grid multilayer inter-DC optical networks: A comparison study," *J. Lightw. Technol.*, vol. 35, pp. 5335–5346, Dec. 2017.