The Scanning the Literature column provides concise summaries of selected papers that have recently been published in the field of networking. Each summary describes the paper's main idea, methodology, and technical contributions. The purpose of the column is to bring the state of the art of networking research to readers of *IEEE Network*. Authors are also welcome to recommend their recently published work to the column, and papers with novel ideas, solid work, and significant contributions to the field are especially appreciated. Authors wishing to have their papers presented in the column should contact the Editor.

Xiaohua Tian, Shanghai Jiao Tong University
xtian@sjtu.edu.cn

With the rapid development of communication networks, users demand better performance with fine-grained QoS, which is challenging for the traditional network framework. Fortunately, artificial intelligence (AI) provides a promising solution to solve these problems. The column in this issue focuses on research about network and AI, particularly deep reinforcement learning for traffic optimization, vehicular networks and 3D wireless cellular networks, machine learning models at the network edge, multi-task learning in radio access networks, and prediction with a probabilistic principal component analysis (PPCA) model.

Traffic optimizations (TO, e.g. flow scheduling, load balancing) in data centers are face difficult online decision making problems. Previously, they have been done with heuristics relying on operators' understanding of the workload and environment. Designing and implementing proper TO algorithms thus takes at least weeks. To solve this problem, Chen *et al.* introduce their parallelism design in the following paper.

AuTO: Scaling Deep Reinforcement Learning for Datacenter-Scale Automatic Traffic Optimization

Li Chen, Justinas Lingys, Kai Chen, and Feng Liu, *Proc. SIGCOMM*, ProgramSunday, 20–25 Aug. 2018.

In this paper, leveraging the long-tail distribution of data center traffic, the authors develop a two-level DRL system, AuTO, mimicking the peripheral and central nervous systems in animals to solve the scalability problem. Peripheral systems (PSs) reside on end hosts, collect flow information, and make TO decisions locally with minimal delay for short flows. PSs' decisions are informed by a central system (CS), where global traffic information is aggregated and processed. CS further makes individual TO decisions for long flows. With CS&PS, AuTO is an end-to-end automatic TO system that can collect network information, learn from past decisions, and perform actions to achieve operator-defined goals. The authors implement AuTO with popular machine learning frameworks and commodity servers, and deploy it on a 32-server testbed. Compared to existing approaches, AuTO reduces the TO turn-around time from weeks to 100 ms while achieving superior performance. For example, it demonstrates up to 48.14 percent reduction in average flow completion time (FCT) over existing solutions.

Emerging technologies and applications including Internet of Things (IoT), social networking, and crowd sourcing generate large amounts of data at the network edge. Machine learning models are often built from collected data to enable the detection, classification, and prediction of future events. Due to bandwidth, storage, and privacy concerns, it is often impractical to send all the data to a centralized location. To fill this gap, Wang *et al.* illustrate their solution in the following paper.

When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning

Shiqiang Wang , Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makay, Ting He, and Kevin Chan, *Proc. INFOCOM*, Honolulu, HI, 16–19 Apr. 2018

In this paper, the authors consider the problem of learning model parameters from data distributed across multiple edge nodes, without sending raw data to a centralized place. Their focus is on a generic class of machine learning models that are trained using gradient-descent-based approaches. They analyze the convergence rate of distributed gradient descent from a theoretical point of view, based on which we propose a control algorithm that determines the best trade-off between local update and global parameter aggregation to minimize the loss function under a given resource budget. The performance of the proposed algorithm is evaluated via extensive experiments with real datasets, both on a networked prototype system and in a larger-scale simulated environment. The experimentation results show that the proposed approach performs near to the optimum with various machine learning models and different data distributions.

Recently, an increasing amount of mobile analytics is performed on data that is procured in a real-time fashion to make real-time decisions. Such tasks include simple reporting on streams to sophisticated model building. However, the practicality of these analyses are impeded in several domains because they face a fundamental trade-off between data collection latency and analysis accuracy. Motivated by this problem, Anand Padmanabha Iyer *et al.* illustrate their idea in the following paper.

Mitigating the Latency-Accuracy Trade-off in Mobile Data Analytics Systems

Anand Padmanabha Iyer, Li Erran Li, and Mosharaf Chowdhury ,Ion Stoica, *Proc. MobiCom*, New Delhi, India, 29 Oct.–02 Nov., 2018

In this paper, the authors first study this trade-off in the context of a specific domain, cellular radio access networks (RANs). They find that the trade-off can be resolved using two broad, general techniques: intelligent data grouping and task formulations that leverage domain characteristics. Based on this, they present CellScope, a system that applies a domain-specific formulation and application of multi-task learning (MTL) to RAN performance analysis. It uses three techniques: feature engineering to transform raw data into effective features, a PCA-inspired similarity metric to group data from geographically nearby base stations sharing performance commonalities, and a hybrid online-offline model for efficient model updates. The evaluation shows that CellScope's accuracy improvements

over direct application of ML range from 2.5× to 4.4× while reducing the model update overhead by up to 4.8×.

Short-term traffic prediction, referring to horizons from a few minutes up to around 60 minutes depending on the network, is a crucial component of intelligent transportation systems (ITS). In the following paper, Erik Jenelius *et al.* propose a network travel time prediction methodology based on probe data, which is intended as a tool for traffic management, trip planning, and online vehicle routing, and is designed to be efficient and scalable in calibration and real-time prediction; flexible to changes in network, data, and model extensions; and robust against noisy and missing data.

Urban Network Travel Time Prediction Based on a Probabilistic Principal Component Analysis Model of Probe Data

Erik Jenelius and Haris N. Koutsopoulos, *IEEE Trans. Intelligent Transportation Systems*, vol. 19, no. 2, June 2018, pp. 436–45

In this paper, a multivariate PPCA model is proposed. Spatio-temporal correlations are inferred from historical data based on MLE and an efficient EM algorithm for handling missing data. Prediction is performed in real time by computing the expected distribution of link travel times in future time intervals, conditional on recent current-day observations. A generalization of the methodology partitions the network and applies a distinct PPCA model to each subnetwork. The methodology is applied to the network of downtown Shenzhen, China, using taxi probe data. The model captures variability over months and weekdays as well as other factors. Prediction with PPCA outperforms *k*-nearest neighbor prediction for horizons from 15 to 45 min, and a hybrid method of PPCA and local smoothing provides the highest accuracy.

The developments of connected vehicles are heavily influenced by information and communications technologies, which have fueled a plethora of innovations in various areas, including networking, caching, and computing. Nevertheless, these important enabling technologies have traditionally been studied separately in the existing works on vehicular networks. He *et al.* show their deep reinforcement learning solution in the following paper.

Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach

Ying He, Nan Zhao, Hongxi Yin, *IEEE Trans. Vehicular Technology*, vol. 67, no. 1, Jan. 2018, pp. 44–55

This paper proposes an integrated framework that can enable dynamic orchestration of networking, caching, and computing resources to improve the performance of next generation vehicular networks. The authors formulate the resource allocation strategy in this framework as a joint optimization problem, where the gains of not only networking but also caching and computing are taken into consideration in the proposed framework. The complexity of the system is very high when they jointly consider these three technologies. Therefore, the authors propose a novel deep reinforcement learning approach in this paper. Simulation results with different system parameters are presented to show the effectiveness of the proposed scheme.

The number of unmanned aerial vehicles (UAVs), also known as drones, will exceed 7 million in 2020. Such massive use of drones will have significant impacts on wireless networking. In such cases, the deployment of aerial drone base stations (BSs) is a promising opportunity for providing reliable wireless connectivity for drone user equipments (UEs). To support drones in wireless networking applications, Mohammad Mozaffari *et al.* introduce the novel concept of a 3D cellular network that incorporates both drone BSs and drone UEs in the following paper.

Beyond 5G With UAVs: Foundations of a 3D Wireless Cellular Network

Mohammad Mozaffari; Ali Taleb Zadeh Kasgari; Walid Saad; Mehdi Bennis; and Mérouane Debbah, *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, Jan. 2019, pp. 357–72.

In this paper, a novel concept of three-dimensional (3D) cellular networks is introduced. For this new 3D cellular architecture, a novel framework for network planning for drone-BSs and latency-minimal cell association for drone-UEs is proposed. For network planning, a tractable method for drone BSs' deployment based on the notion of truncated octahedron shapes is proposed. In addition, to characterize frequency planning in such 3D wireless networks, an analytical expression for the feasible integer frequency reuse factors is derived. Subsequently, an optimal 3D cell association scheme is developed for which the drone UEs' latency, considering transmission, computation, and backhaul delays, is minimized. To this end, first, the spatial distribution of the drone UEs is estimated using a kernel density estimation method, and the parameters of the estimator are obtained using a cross-validation method. Then, according to the spatial distribution of drone UEs and the locations of drone BSs, the latency-minimal 3D cell association for drone UEs is derived by exploiting tools from an optimal transport theory. The simulation results show that the proposed approach reduces the latency of drone UEs compared to the classical cell association approach that uses a signal-to-interference-plus-noise ratio criterion.