A4E: WHEN AI MEETS WIRELESS

The Emerging Role of AI/ML in Next-Generation Networks Panel

IEEE INFOCOM 2022 Workshop—A4E: AI/ML for Edge/Fog Networks
May 2, 2022

Shiwen Mao
Auburn University, Auburn, AL, USA
**Intelligence**: will be the definitive feature of the NextG
Wireless communications and networks:
- Well studied with good understanding, lot of models and theory
- But systems get more and more complicated, heterogenous, large-scale, and dynamic

AI/ML:
- Huge success in other fields: natural language processing, computer vision, gaming, ... → can be replicated in NextG
- Epic events (good advertisement):
  - IBM Deep Blue vs. World Chess Champion Garry Kasparov (1997)
  - IBM’s DeepQA project: quiz show Jeopardy! won the first-place prize (2011)
  - Google DeepMind’s AlphaGo/AlphaGo Zero: beat Ke Jie, the world No.1 ranked Go player (2017)
  - Facebook/CMU’s Pluribus: beat 15 of the world’s top poker players (2019)
  - Dr. Fill, Champion of the 43rd Annual American Crossword Puzzle Tournament (2021)
  - DeepMind: solved a 50-year-old challenge in biology, and helped crack two mathematical puzzles that stumped humans for decades (2021)
  - ...
- Technology is ready:
  - Availability of: Data, Computing power, and open-source Platforms
  - Wireless designs: historically based on probabilistic models (e.g., traffic, channel, interference, ...), and are fault tolerant

Challenges: data, generalization, explainable
Case Study: Edge Computing/Offloading

• Considers:
  • Local execution, and offloading (to which BS)
  • CPU frequency tuning
  • Energy harvesting
  • Mobility/handover
• Control knobs ($c$, $e$):
  • $c$: offloading or local execution
  • $e$: energy allocation
• A Markov Decision Process: to maximize the long-term utility by offloading decisions made with the current system state: task queue state, energy queue state, channel quality
  • A double deep Q-network (DQN) algorithm to learn the optimal offloading policy
  • Exploiting the additive structure of the utility function, decompose the Q-function and use a “virtual” agent to learn each of the decomposed element

Traditional analytical methods is not capable of handling such formulated problems

Case Study: RF Sensing/3D Human Pose Tracking

Wireless Engineering Research and Education Center

Wireless for AI

Motivation:
- ML usually involves a large amount of data, large-sized models, high computing requirements, and stringent delay requirements
  - For example: autonomous driving
    - One autonomous car will use 4,000 GB of data/day [2]
    - 70 miles/hour ⇒ 31.3 m/s: decision should be made in ms or 10s of ms
- Limited computing power and storage at mobile devices
- Privacy concerns: not willing to share data

Solutions:
- Distributed machine learning over wireless networks
  - Federated Learning
  - Vertical or horizontally partition a large ML model
- Fundamentally new wireless communication protocols to support future machine learning applications

Benefits:
- Improvement in energy efficiency
- Improvement in bandwidth utilization
- Improvement in latency
- Training ML model in a privacy-preserving manner

Challenges
- Communication efficiency
- System heterogeneity
- Statistical heterogeneity, non-i.i.d. data
- Privacy preservation, comprised devices, polluted data

Case Study: IRS-assisted Federated Learning

**Problem:** Intelligent reflecting surface (IRS) assisted federated learning for higher energy efficiency

- Federated learning is a distributed learning algorithm for better protecting user data privacy, with many applications.
- IRS: regarded as a candidate technology for 6G, can create a smart radio environment.
- Wireless for Machine Learning: how to use IRS to make more efficient federate learning.

**Main results:**

- Joint beamforming and resource allocation problem, aiming to optimize the energy efficiency under federated learning’s training time constraint.
- Developed an alternative optimization approach, which decompose the original problem into multiple semidefinite programming (SDP) problems.
- Leveraged the majorization-minimization (MM) algorithm and developed a low-complexity algorithm with greatly reduced complexity with similar optimality performance.
- Considerable energy savings found in our simulation study.

---

Three levels of problems of communications (Shannon and Weaver)

- Technical – How a channel causes a problem
- Semantic – Is the meaning of message sent and received very different
- Influential – How effectively does the message cause reaction by the receiver

Semantic communication: the shared and local knowledge, extracted from the set of transmission contents, helps compress transmission information and correct the transmission errors according to semantic correlation.

Fig. A comparison between (a) tradition communication and (b) semantic communication (This figure is from [1])

Benefits:

- Consumes less radio resources
- Reduced latency
- Relatively insensitive to channel noise, improved transmission reliability
- Potentially avoid the co-channel interference
- Integrated with DNN well, enable intelligent end-to-end communication

Challenges

- Insufficient theory
- Performance metrics, design trade-offs
- Inconsistent knowledge base (KB) between the source and destination
- Semantic resource allocation
- Practical implementation

Applications:

- Self-driving
- IoT smart home
- Virtual reality
- Industrial IoT
- Photo
- Text
- Video
- Speech

S. Mao’s work is supported in part by the National Science Foundation (NSF) under Grants ECCS-1923163, ECCS-1923717, CNS-2107190, and CNS-2148382, and through the Wireless Engineering Research and Education Center at Auburn University.

For more information: http://www.eng.auburn.edu/~szm0001/