A Survey on the Role of Artificial Intelligence in Mobile Systems

Challenges and Opportunities

Assignment 1
COMP4342 Mobile Computing

MA Mingyu Derek, 14110562D
Department of Computing
The Hong Kong Polytechnic University
derek.ma@connect.polyu.hk
https://derek.ma

March 9, 2018

Abstract

Complex tasks for Mobile Systems need to be handled by intelligent high-performing algorithms. Artificial Intelligence is renovating the industries and it needs an entry for personalized data and mobile scenarios. There are some existing works for combining AI capability with Mobile Systems. The agent-based architecture provides fundamental structure to deploy AI solutions on the mobile devices. Natural Language Understanding helps the Human-Computer Interactions on mobile devices and Computer Vision technology provides an intelligent classifier and image enhancement ability to the mobile systems.

There are many challenges for applying current AI techniques into mobile systems nowadays. Limitation of computational power on mobile systems limits the scale and performance of AI algorithms; poor utilization of distributed resources constrains the coordination between agents; poor environment detection capability makes the mobile devices hard to accommodate the environment; potential risk on privacy protection for personal data may stop people and organizations utilizing AI on mobile systems.

In the meantime, some clear directions and opportunities corresponding to the challenges may bring mobile systems into next level. Hardware acceleration for deep learning is a chance to resolve the limitation of computational power and deploy AI algorithms in a larger scale; multi-agents distributed algorithms and networking can promote coordination and collaboration between agents to collect more data and computational power to accomplish heavy tasks; advanced context awareness solution is a great opportunity to customized behaviour of mobile systems based on AI techniques; customization of content and development of cryptography is an important occasion to personalized user experience on mobile systems and keep users anonymous to protect their privacy.

We states that AI empowers new applications and capabilities for next-generation mobile systems.

Contents

1 Introduction

With the rapid increment of hardware performance, *Artificial Intelligence*, a technology that has been researched for several decades, is leading the new trend of computing evolution. Data collected from several types of mobile devices becomes a foundational resource for AI techniques, while AI techniques can also help to improve the performance of mobile systems and empower new capability for mobile systems (LeCun et al., 2015).

As the performance and ability of mobile devices are increasing dramatically, mobile devices are becoming smaller in term of size and closer in term of relationships with the users. For example, billions of people stores their sensitive personal behaviour data and personal photos, messages on smartphones (Ebrahimzadeh and Maier, 2016). This makes mobile systems which consist of many mobile devices the great entry for personal data which is wistful and useful in AI algorithms especially *Deep Learning* algorithms.

AI can also create new applications of mobile systems. Advanced computing algorithms and smarter architecture are helping the mobile systems become smarter for individuals and more professional for businesses (Purdy, 2016). AI can help to allocate resources like data and bandwidth among mobile devices with higher efficiency. What's more, the capability of accomplishing more complex calculation such as path planning, image processing provided by AI can also let mobile systems be deployed to new problems and scenarios to solve more difficult tasks.

The role of AI in the mobile systems would be illustrated like this: **Artificial Intelligence empowers new applications and capability for next-generation mobile systems** (Krüger and Malaka, 2004).

The other parts of the paper are organized like the following. Existing works of the combination of Artificial Intelligence and Mobile Systems will be introduced in Section 2, and major challenges for AI in mobile systems will be argued in Section 3 where four challenges are stated. Corresponding to the challenges, four opportunities will be explained in Section 4. Finally, the conclusion is made in final Section 5.

2 Existing Works

2.1 Deployment of AI on Agent-based Architecture of Mobile Systems

Agent-based architecture is utilizing every agent to responsible part of the whole task (Aversa et al., 2010). In the contrast, this architecture does not use centralized computers to calculate the results. Existing works on the deployment of AI are normally using this architecture (Shen and Norrie, 1999). There is also some AI framework that already supports agent-based architecture such as *TensorFlow Lite*. Previously, people deployed AI algorithms like recommendation algorithms or Natural Language Processing algorithm on the cloud, and then transfer the calculation result to the mobile systems through networks. But with agent-based architecture, AI algorithms can be deployed on the mobile devices. Some works already built good-performing *Optical Character Recognition (OCR)* classifier locally on the mobile devices and some researchers use a cluster of mobile devices to run AI algorithms (Jayadevan et al., 2011).

2.2 Natural Language Understanding

Since the interface for user interaction on mobile devices are normally not large, and mobile systems lack assistant input devices like mouse and keyboard, an efficient way to communicate between the user and the mobile system is very important. Natural Language Understanding (NLU) is a group of technologies to convert users' spoken language to formal input to the computer system, it can also generate natural language response to the users (Huang et al., 2001). Mobile systems is a great showcase for the NLU techniques and it does solve interaction problems for mobile systems. NLU techniques are deployed on personal mobile devices like smartphones for years with the format of voice assistant and voice input engine. *Google Assistant*, *Apple Siri* and *Microsoft Cortana* are widely used voice assistants. NLU is also used in professional areas like controlling robots (Tellex et al., 2011).

2.3 Computer Vision

Computer Vision techniques utilize multi-layers neural networks to process tons of images to build an image classifier and achieve detailed modification for images. Mobile devices like robots or smartphones collect a large number of images or videos, and Computer Vision and utilize these data (Fei-Fei and Perona, 2005). Computer Vision classifier on mobile systems

can be used to recognize faces, scenes, generate videos and do automatic danger screening. Advanced computer vision technique can adjust special effect, background automatically. Computer Vision technology is widely used in album app(Google Photos), security authentication(Apple FaceID), and camera image enhancement etc.

3 Challenges

Applying Artificial Intelligence in mobile systems is still facing some critical challenges in terms of hardware capacity, software algorithms and even law and social regulations (Purdy, 2016). In the following part of this section, some detailed statements of current major challenges will be made. They are:

- Limitation of Computational Power on Mobile Systems
- Bad Utilization of Distributed Resources
- Poor Environment Detection Capability
- Potential Risk on Privacy Protection for Personal Data

3.1 Limitation of Computational Power on Mobile Systems

Current approaches and applications of Artificial Intelligence heavily rely on *Deep Learning (DL)* algorithms as stated by LeCun et al. (2015). DL is a machine learning techniques which build neural networks with many layers and many computational hops. Input data are fed into the network to train the parameters inside the network to achieve an optimized network performance. (LeCun et al., 2015)

Normally, industry-level DL model are trained using a cluster of *Graphics Processing Unit(GPU)* or even *Tensor Processing Unit(TPU)* which are expensive high-performance application-dependent hardware devices (Schmidhuber, 2015). Even the CPU performance on smartphones is increasing rapidly, but there is still a huge gap of computational capacity between GPU/TPU and CPU on the phones in term of the order of magnitude. A comparison of CPU, GPU and TPU capacity is shown in Figure 1 (Sato et al., 2017).

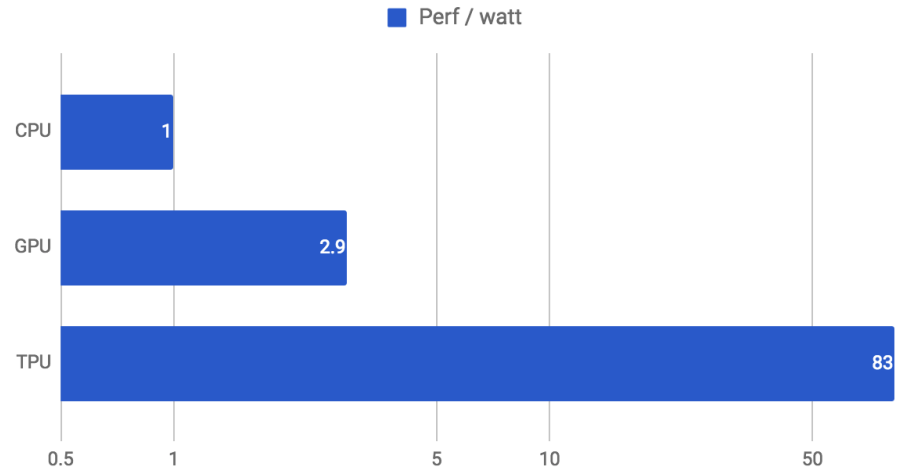

Figure 1: Capacity comparison among CPU, GPU and TPU by Sato et al. (2017)

Due to the computational power constraint, there are few instances that use mobile systems as the computational nodes for AI algorithms. Most of the current solution is to train the AI model on cloud and then use network connect the model with the mobile systems (Song et al., 2014). So it's hard to fully deploy an AI algorithm on mobile systems because the mobile systems still need the network connection and the support of cloud computational resources.

3.2 Bad Utilization of Distributed Resources

Current mobile systems like smartphones, multi-agents *Internet of Things(IoT)* devices, multi-robots systems are consists of many single instances in the system. As stated in Section 3.1, every single instance is equipped with very limited action capability and computational capacity. But actually, the mobile systems with many instances can provide considerable action capability and computational power, but the design of a general mechanism for the collaboration and coordination between instances in the system is still a big challenge (Jiang et al., 2016).

3.3 Poor Environment Detection Capability

A very important precondition for utilization of AI algorithm is that a large amount of data from different perspectives and sources need to be provided.

For example, to recommend shopping items for users, provided data may include: location, time, income level, gender, past shopping history, neighborhood, weather etc. Current instances of mobile systems are normally equipped a few sensors like distance sensor, light sensor and digital camera. But this data are not sufficient for the AI algorithms to utilize because the data are within limited range and limited perspectives. The mobile device can't obtain information beyond its capability, and it cannot utilize the data from other different type mobile devices in the same scene (Chen et al., 2000). Poor environment detection capability becomes a constraint for applications of AI in mobile systems (Abowd et al., 1999).

3.4 Potential Risk on Privacy Protection for Personal Data

Boyd and Crawford (2012) states the concern of privacy protection in big data era. Mobile systems especially mobile phones are with the owner every day. Through mobile systems, sensitive personal behaviour data can be collected (Krüger and Malaka, 2004). According to current deployment of AI techniques in mobile systems, there is no clear law or social regulations to constrain the collections and utilization of collected of personal data. Business giants in the industry have their different approaches to handle sensitive personal privacy. Google, Facebook and Microsoft will hash the user identity and collect back to the Cloud, while Apple utilizes a technique called *differential privacy* to add random data into the privacy data before transferring data back to the server. Data is the fundamental input resources to AI algorithms, while privacy is the biggest concern for collecting data. The potential risk for privacy data collection may stop the community from utilizing AI algorithms in very private mobile systems like smartphones and that is also a very big challenge for the AI and Mobile System communities.

4 Opportunities

As the value of the application of AI in mobile systems is very huge in terms of business and academic, there are some clear directions and opportunities which is urgent needs of the industry and potentially valuable research topics. They are corresponding to the issues, problems and challenges put forwarded in Section 3 and targetting solve these issues to promote the applications migration of Artificial Intelligence technologies in mobile systems.

Corresponding to the four challenges in Section 3, some potential directions or opportunities are:

- Hardware Acceleration for Deep Learning (to deal with Challenge 3.1)
- Multi-agents Distributed Algorithms and Netowrking (to deal with Challenge 3.2)
- Context Awareness Solution (to deal with Challenge 3.3)
- Customization of Content and Behaviour and Development of Cryptography (to deal with Challenge 3.4)

4.1 Hardware Acceleration for Deep Learning

To enable the application of AI techniques in mobile systems especially Deep Learning algorithms, some specific hardware chips and components can be designed to accelerate the calculation for Deep Learning operations. These *Acceleration Hardware* is designed to benefit limited operations which is very common in Deep Learning such as *Matrix Multiplication*, *Convolution*, Calculation of Recurrent Layers and so on (Lacey et al., 2016).

Besides NVIDIA's Deep Learning GPU and its programming framework *CUDA*, there are some new competitors in this market. *Xeon Phi Knight's Landing* by Intel, *Neural Network Processor (NNU)* by Qualcomm (Ross et al., 2017) and *Field-programmable Gate Array (FPGA)* by Teradeep are all acceleration hardware for deep learning (Wang et al., 2017a). *Performance per Watt* is the index to measure the performance of hardware architectures. These acceleration hardware can achieve 10-100 times faster processing speed in terms of AI algorithms compared to traditional CPU.

In the future, many smartphones and other mobile systems will be equipped with Deep Learning Acceleration Hardware to speed up local on-machine AI calculation. There is still a huge potential research blank and market space in this field.

4.2 Multi-agents Distributed Algorithms and Networking

A mobile system consists of many agents. A mobile network connects many phones, computers and laptops. A mobile robotic system may contain many small robotics. Collaboration and coordination between multi-agents can

enable every single agent to achieve more using AI technology. Provchain proposed by Liang et al. (2017) can use blockchain to collaborate between single computers or phones to do data storage. A distributed algorithm for multiple robots proposed by Wang et al. (2017b) can collaborate them to behave together.

Collaborating multiple agents can gather the computational power of all agents and share the computational power together. A more intelligent computational power sharing mechanism can be designed to distribute the joined computer power according to tasks priority and reduce repeated computation. Coordinating multiple agents can build an advanced *Content Delivery Network (CDN)* to share downloaded data and utilize joined data train the joined model or use joined computational power in a better way. For example, benefit from AI techniques in path planning, now a cluster of drones can connect together and share their computational power to calculate best flight path and coordinate their behaviour to achieve a task.

With more advanced multi-agents distributed algorithms and networking, more new applications of AI technology in Mobile Systems can be developed to increase the productivity in many industries. So it's a very important opportunity in the AI and big data era.

4.3 Context Awareness Solution

As stated in Section 3.3, environment detection can provide the mobile device background information about its situation and create more personalized and customized application of AI from the very end of the users. With more sensors on mobile devices and more advanced AI technology, context awareness computing is a crucial opportunity.

A basic context awareness method in mobile devices is proposed by Gellersen et al. (2002). For example, previously, if a mobile device would like to aware the context in the room, it needs to ask for more data like the location of the room, the current schedule, weather information and so on. But with AI technology especially Computer Vision techniques, it only needs to capture the scene using digital camera, the Computer Vision algorithm can identify the context in the room.

With better context awareness solution, mobile systems can act with personalization and customization. The user experience of mobile devices like smartphones would be much better (Chen et al., 2000).

4.4 Customization of Content and Behaviour and Development of Cryptography

Modern mobile devices are becoming *disappearing computers*, which they are becoming smaller and saving much sensitive individual information. People are spending lots of time on mobile devices and will spend more. Such a personal device is a very good entry for personalized and customized content delivery (Gupta, 2017).

The next generation applications will combine more intuitive interaction ways and reach users in a deeper way. By leverage AI and data collected from the mobile device, personalized software and content is much easier to product (Dossey, 2017). Internet giants like Amazon and Google are starting launching products and services that utilizing personal data from mobile systems, *Amazon Alexa* and *Google Home* are such outputs. How to utilize personal data from mobile systems and create advanced customized and personalized content and experience would be a great opportunity in terms of combining AI and mobile systems.

While privacy protection is still a concern as introduced in Section 3.4, *Cryptography* is a great tool to deal with the privacy issue in the big data era. Chan and Blake (2005) proposed a cryptography approach to keep user anonymous when the personal data is collecting and another work by Huang (2007) also aims at this target. Differential Privacy algorithm used by Apple Inc. is also a good application of cryptography to deal with privacy protection in mobile systems (McSherry and Mironov, 2013).

5 Conclusions

Mobile system is a data collection entry for Artificial Intelligence algorithms, while AI algorithms empower next generation applications and capability for the mobile systems. Existing works on Agent-based Architecture, Natural Language Understanding and Computer Vision have already been deployed in industry-level products.

There are still many challenges in terms of applying AI in mobile systems, such as limitation of computational power, bad utilization of distributed resources, poor environment detection capability and potential risk on privacy protection for personal data. But corresponding to those challenges, there are some directions and opportunities. Hardware Acceleration for Deep Learning

can be used to resolve the limitation of computational power, Multi-agents Distributed Algorithms and Networking can be deployed to utilize distributed resources, Context Awareness Solution can improve the environment detection capability and finally Customization of Content and Behaviour and Development of Cryptography can utilize personal data to improve user experience while protecting personal data.

References

Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer, 1999.

Rocco Aversa, Beniamino Di Martino, Massimiliano Rak, and Salvatore Venticinque. Cloud agency: A mobile agent based cloud system. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on*, pages 132–137. IEEE, 2010.

Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

AC-F Chan and Ian F Blake. Scalable, server-passive, user-anonymous timed release cryptography. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 504–513. IEEE, 2005.

Guanling Chen, David Kotz, et al. A survey of context-aware mobile computing research. Technical report, Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College, 2000.

Annie Dossey. How artificial intelligence is driving mobile app personalization, Oct 2017. URL https://clearbridgemobile.com/artificial-intelligence-driving-future-mobile-app-personalization/.

Amin Ebrahimzadeh and Martin Maier. Artificial intelligence based mobile-edge computing, 2016. URL http://zeitgeistlab.ca/doc/artificial_intelligence_based_mobile-edge_computing.html.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

Hans W Gellersen, Albercht Schmidt, and Michael Beigl. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mobile Networks and Applications*, 7(5):341–351, 2002.

Robby Gupta. How businesses can use artificial intelligence in mobile apps, Apr 2017. URL http://www.techjini.com/blog/businesses-can-use-artificial-intelligence-mobile-apps/.

Dijiang Huang. Pseudonym-based cryptography for anonymous communications in mobile ad hoc networks. *International Journal of Security and Networks*, 2(3-4):272–283, 2007.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*, volume 95. Prentice hall PTR Upper Saddle River, 2001.

R Jayadevan, Satish R Kolhe, Pradeep M Patil, and Umapada Pal. Offline recognition of devanagari script: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):782–796, 2011.

Shan Jiang, Jiannong Cao, Yang Liu, Jinlin Chen, and Xuefeng Liu. Programming large-scale multi-robot system with timing constraints. In *Computer Communication and Networks (ICCCN), 2016 25th International Conference on*, pages 1–9. IEEE, 2016.

Antonio Krüger and Rainer Malaka. Artificial intelligence goes mobile. *Applied Artificial Intelligence*, 2004.

Griffin Lacey, Graham W Taylor, and Shawki Areibi. Deep learning on fpgas: Past, present, and future. *arXiv preprint arXiv:1602.04283*, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

Xueping Liang, Sachin Shetty, Deepak Tosh, Charles Kamhoua, Kevin Kwiat, and Laurent Njilla. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 468–477. IEEE Press, 2017.

Frank D McSherry and Ilya Mironov. Differential privacy preserving recommendation, December 31 2013. US Patent 8,619,984.

J Gerry Purdy. Why artificial intelligence is important to mobile, Jul 2016. URL https://www.rcrwireless.com/20160713/analyst-angle/analyst-angle-artificial-intelligence-important-mobile-tag9.

Jonathan Ross, Norman Paul Jouppi, Andrew Everett Phelps, Reginald Clifford Young, Thomas Norrie, Gregory Michael Thorson, and Dan Luu. Neural network processor, July 18 2017. US Patent 9,710,748.

Kat Sato, Cliff Young, and David Patterson. An in-depth look at google's first tensor processing unit (tpu), May 2017. URL https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Weiming Shen and Douglas H Norrie. Agent-based systems for intelligent manufacturing: a state-of-the-art survey. *Knowledge and information systems*, 1(2):129–156, 1999.

Inchul Song, Hyun-Jun Kim, and Paul Barom Jeon. Deep learning for real-time robust facial expression recognition on a smartphone. In *Consumer Electronics (ICCE), 2014 IEEE International Conference on*, pages 564–567. IEEE, 2014.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, page 2, 2011.

Chao Wang, Lei Gong, Qi Yu, Xi Li, Yuan Xie, and Xuehai Zhou. Dlau: A scalable deep learning accelerator unit on fpga. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 36(3):513–517, 2017a.

Jia Wang, Jiannong Cao, and Shan Jiang. Fault-tolerant pattern formation by multiple robots: a learning approach. In *Reliable Distributed Systems (SRDS), 2017 IEEE 36th Symposium on*, pages 268–269. IEEE, 2017b.