A Survey of Deployment of Artificial Intelligence

Kaixin Wu

wkx2562973335@gmail.com

School of Computer Engineering Guangzhou City University of Technology Guangzhou, China

Abstract—The development of artificial intelligence systems centered on machine learning has become a new trend in the development of world science and technology. With the research of more and larger enterprises, the artificial intelligence system has become perfect. In the process of system improvement, the deployment of artificial intelligence system has also become a challenge for the academic community. How to choose the best architecture or system to deploy artificial intelligence system is the challenge of this research. This paper studies the deployment environment and platform of artificial intelligence system. Distinguish between many different types of architectures and platforms. It also explains the best choice of different architectures or platforms in the environment. The realization of intelligent system is analyzed, and ration and quantitative machine learning algorithms are used to solve the system construction. Summarize the calculation methods of different types of intelligent systems and some problems encountered. In addition, on this basis, through the deployment and evaluation of different artificial intelligence systems, this paper found many development fields that may be involved in the future.

Keywords-artificial intelligence system, machine learning, deployment environment, qualitative, ration

1 INTRODUCTION

With the development of artificial intelligence, more and more large enterprises need to use artificial intelligence systems. In response to the new trend of world science and technology development, this paper discusses the key factors of AI system deployment and the choice of AI deployment architecture. Taking systems with different environmental requirements as examples, this paper seeks the most suitable scheme for the research of artificial intelligence system deployment.

The second chapter of this paper, as the classification of research problems, describes the deployment platform used in two major fields, industrial and commercial environment, as the classification conditions. Deployment platform is divided into architecture and system. The deployment of artificial intelligence system through these two platforms is the characteristic of this research. It can intuitively express the performance impact of different architectures or systems on artificial intelligence in different environments.

ITEI 2022, December 23-25, Harbin, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.23-12-2022.2329084

The third part of this paper is the classification of research methods. In the process of deploying artificial intelligence system, the implementation method is the key factor. Because of researchers' interviews and questionnaires, they use qualitative methods to analyze the collected data. Under the qualitative research method, researchers use subjective and objective realization to study the problem. Use quantitative means to process digital data.

The fourth part is a summary of experimental analysis. In the experimental analysis, we found out OE, CTR and other indicators and corresponding parameters. Then the indicators are evaluated. Each index has different meanings for different research problems. After evaluating the indicators, we came up with an indicator formula. Finally, we explained the significance of all the parameters, and compared different experiments.

The fifth part is discussion on the research methods and research objects of various references, as well as some innovative research on artificial intelligence systems, including the application of RNN-based algorithm and ML technology in developing intelligent systems. There is also the use of modeled machine learning models to detect session-based fraud detection in online e-commerce transactions, and so on.

The sixth part is the summary. It mainly describes research field and research methods. Finally, it summarizes some difficulties encountered in this research, as well as some challenges and opportunities in the future.

The rest of this article is organized as follows. The second part gives a classification of research objects for the problem of artificial intelligence deployment. The third part introduces the classification of research methods. The fourth part introduces the comparison of experimental analysis in the relevant literature. Section V discusses research opportunities in future work, and Section VI summarizes the paper.

2 CLASSIFICATION OF RESEARCH OBJECTS

2.1 Criteria

With the in-depth study on the deployment of artificial intelligence and the investigation of research objects, we have found many research topics that can be classified. In this section, two independent and different criteria would be used to divide research objects into different types:

1) **Deployment environment**. In this paper, we divided the deployment environment of a theme into two parts. One is the commercial environment, and the other is an industrial environment. Actually, the evolution of machine learning and the deployment of systems can be embodied in many environments, among which the business environment is closely related to our real lives. Therefore, artificial intelligence in the commercial environment has become one of our classifications. Similarly, the industry environment is also the object of our classification.

2) **Platform**. There are two kinds of platform here: architecture or system in the platform. The architecture is the foundational module of artificial intelligence, and its role in the environment we deploy is extremely important. Because of different deployment environments, the required

architecture is naturally different. On the other hand, as the core of artificial intelligence, the system is a runtime platform for applications, which can be seen as implementations of various architectures.

2.2 The Classification

Based on the appeal classification standard, we give the classification in Table 1. The meaning of each class is as follows:

Type I: This type is a combination of architecture and commerce of AI deployment systems.

Type II: This type is a combination of systems and commerce for AI deployment systems.

Type III: This type is a combination of architecture and industry of AI deployment systems.

Type IV: This type is a combination of systems and industry for AI deployment systems.

Table 1: Different Research Objects

Deployment environment	**Platform**	
	Architecture	**System**
Commerce	I. [1]	II. [3][4] [5]
Industry	III. [2]	IV. [6]

2.3 Explanation of Different Types

References ([1]) belong to Type I. It lists many different types of AI deployment architectures. In addition, through experimental comparison, the one with the best performance is selected.

References ([3] [4] [5]) belong to Type II. Reference [3] suggests the commercial value of machine learning. These include large-scale research on the impact of machine learning in the field of commercial products for the first time and a collection of "lessons learned" covering all phases of a machine learning project. The result is a set of technologies that solve the challenges we find at each stage of the project. Reference [4] raises the question of how to build large-scale product recommendation systems. (Sigmund system eventually became the most popular product recommendation system). Reference [5] uses RNN-based algorithms to detect session-based fraud in recurrent neural network online e-commerce transactions. The RNN-based algorithm can reflect the detection performance of the structure, and a more suitable detection structure can be selected through comparison. In the α-β structure, the number of layers and units is also a criterion for the performance of the algorithm.

References [1] belong to Type III. It presents a nine-phase workflow process for AI application development, a set of best practices for building applications and platforms based on machine learning, a customized machine learning process maturity model to assess software teams' progress in building AI applications. It also discusses the three fundamental differences between how software engineering can be applied to machine learning-centric components and previous application areas.

References ([5]) belong to Type IV. It explains how to develop intelligent systems using machine learning (ML) techniques in large enterprise environments

3 CLASSIFICATION OF RESEARCH METHODS

3.1 Criteria

In this section, two independent and different criteria would be used to divide research methods into different types:

1) **Implement**. There are two types of implementations of the research methods. One is subjective and the other is objective. Subjectivity is a non-quantitative way of presenting the thoughts of the experimenter during the experiment and getting the experimental results. Objective is the experimental result presented after calculation by algorithm through quantitative means.

2) **Machine learning**. There are two kinds of manners when applying machine learning to arrive at experimental results: qualitative or ration. The quantitative method can Specific analysis of experimental data derived using different algorithms make the experimental results more logical, theoretical, and more accurate. However, in some aspects of artificial intelligence, quantitative methods cannot fully represent our experimental results, so we use qualitative methods, i.e., rational methods. Combined with the personnel engaged in related professions, visits, and investigations, a second experimental result presented from a different perspective is obtained.

3.2 B The Classification

Based on the appeal classification standard, we give the classification in Table 2. The meaning of each class is as follows:

Type I: This type is an experimental method that is subjective from a qualitative point of view.

Type II: This type is an experimental method of subjective from a rational point of view.

Type III: This type is an experimental method that is objective from qualitative point of view.

Type IV: This type is an experimental method that is objective from a rational point of view.

Table 2. Different Research Methods

Implement	**Machine learning**	
	Qualitative	**Rational**
Subjective	I. [2][3][4]	II.
Objective	III. [1][6]	IV. [5]

3.3 Explanation of Different Types

References ([2] [3] [4]) belong to Type I. In reference [2], it shows a nine-stage workflow for AI application development and a set of best practices for building machine learning-based applications and platforms are addressed by interviewing experiment stakeholders and sending out questionnaires on AI and ML topics. Reference [3] builds a machine learning model and enumerates other members of the machine learning model family. The model will go through five stages: initialization, modeling, deployment, monitoring, and evaluation. Reference [4] builds a factorization model and chooses a factorization-based model, which has been shown to work effectively. Training a separate factorization model for each retailer can increase the amount of data for the experiment and make the results more accurate.

References ([1] [6]) belong to Type III. Reference [1] uses a qualitative research approach and many architectures were sought. Among them, it is roughly divided into global optimum and local optimum. Reference [6] provides qualitative analysis of the interviewed data. Through the interview to different data types, use percentages to represent the level of hierarchy.

References ([5]) belong to Type IV. It uses an RNN-based algorithm. An RNN is a model that captures a range of operations. RNN has been successfully applied in the fields of language translation and speech recognition. N has a recursive structure across time domain, so it can "remember" the information in previous actions and bring this "memory" into the current learning process. At the same time, the parameters of RNN are shared in different time slots, so that it can handle sequential input with variable size.

4 REVIEW OF EXPERIMENTAL ANALYSIS

To ensure the accuracy of the experiment, we measure the experimental results in a qualitative and quantitative manner. However, during the experimental process, the qualitative experimental results will be added to the interviews of various experimental personnel. Therefore, under the premise of realizing the experimental results, we added objective and subjective views to the experiment. In this section, we will classify the metric of evaluation and system parameters, as shown in Table 3. In Table 3, all experimental analysis is also classified according to the metric and parameters. It can be seen from Table 3 that most of the references compare *Precision*, *OE*, *CTR*, and *Participant validation levels*.

Table 3. Experiments with Different Metric and Parameters

Metric	**Parameters**						
	Recall	**Layer/Unit**	**Time period**	**Domain**	**Popularity**	**Issues**	**Phenomenon**
Precision	[5]	[5]	[5]				
OE				[2]			
CTR					[4]		
Participant validation levels						[6]	[6]

4.1 A Metric of Evaluation

Precision means that the threshold test uses the recall rate as the horizontal axis and the precision as the vertical axis to reflect the performance of the α-β RNN structure. The formula is as follows:

$$\mathrm{Precision} = \mathrm{Recall} * \frac{\mathrm{Layer}}{\mathrm{Unit}}$$

Here, *Recall* stands for the recall rate, *Layer* stands for the number of layers of the RNN algorithm structure, and *Unit* stands for a unit of the RNN algorithm structure.

Another kind of the threshold test is based on the retroactive horizontal axis and the accuracy of the vertical axis to reflect the performance of the α-β RNN structure. The formula is:

$$\text{Precision} = \text{Recall } (30\%) * (time\ period\ [a] + time\ period\ [b])$$

The above equation is the definition of the accuracy of the RNN structure at different layers/units at a fixed recall rate of 30%. Here, *Recall (30%)* means precision of different RNN structures under the recall of 30%, *time period [a]* means the precision for the next time period, and *time period [b]* means the precision for the last time period.

OE is average *Overall Effectiveness* for experimental data. The formula is as follows:

$$\text{OE} = \frac{\text{Domain}}{ANOVA\ and\ Knott.test}$$

Here, *Domain* is divided by application field, *ANOVA and Knott.test* is a test method.

CTR means the average *click-through rate* of the product. The formula is as follows:

$$\text{CTR} = \frac{\text{days}}{Daily\ page\ views}$$

In the above equation, *days* means the provisions of the time frame, and *Daily page views* means the browsing of individual pages within the time frame.

Other metric includes *Participant validation levels* etc.

4.2 B System Parameters

Recall represents Different recall rates represent different recall models. The accuracy of a particular recall model is obtained by threshold testing. Evaluate the most appropriate RNN model using the (P-R) curve.

Layer/Unit represents α-βRNN to denote a RNN structure with αlayers and βhidden units per layer.

Time period represents two time periods of the same length. Experiments were conducted with two sets of four consecutive isosceles periods, and both sets showed the effectiveness of the model correction method.

Domain represents nine of the most representative AI application areas. The aberration analysis and Scott Knot's test in nine sets of experiments showed significant differences in reported values, demonstrating the potential value of the indicator in identifying maturity levels in various ML processes.

Popularity represents average popularity within 7 days of all retailers served. Popularity is measured by the number of times the item is displayed each day.

Issues represent crosscutting questions in interviews.

Phenomenon represents phenomena in interviews.

4.3 C Experimental Comparison

In reference [1], the author compares five different deployment architectures of artificial intelligence. It also describes the priorities and trade-offs among the crucial factors of AI deployment, and how to choose a specific architecture. Besides the key factors, the author also discusses other factors that may or may not affect the choice of the best architecture.

In reference [2], the author compares description of nine-stage workflow of how several software engineering teams integrate machine learning into application and platform development. It also puts forward a set of best practices for building applications and platforms based on machine learning and discuss how software engineering is applied to components with machine learning as the core, and three basic differences between them and previous application fields. At the same time, a customized maturity model of machine learning process is made to evaluate the progress of software teams in building artificial intelligence applications.

In reference [3], the author compares six lessons learned from the successful development of 150 large-scale e-commerce applications. At the same time, it is the first time in this field to conduct large-scale research on the influence of machine learning on commercial products. A collection of experimental questions covers machine learning projects. In addition, describes a set of techniques for finding challenges in each project stage.

In reference [4], the author compares two product recommendation services. In addition, the author also describes the basic principles of alternatives considered in various design decisions. At the same time, it will stimulate future academic research in this field and provide information for practitioners who use machine learning to build large-scale services.

In reference [5], the author compares several different RNN structures. In addition, the algorithm based on RNN is used to capture fraud detection in e-commerce websites. After that, it is optimized for the uncertain concept drift. At the same time, it went online to JD.com and deployed CLUE to realize the real-time detection of fraudulent transactions.

In reference [6], the author compares interview data from 11 people who participated in different intelligent system development groups. In addition, qualitative analysis of the data, describes some unique problems and difficulties. At the same time, the source of these unique difficulties and the possible development direction in the future are considered.

5 DISCUSSION AND SUGGESTION

This paper discusses the research methods and research objects of various references and finds that Innovative research on application programs and machine learning algorithms of artificial intelligence system with machine learning as the core. Therefore, this paper puts forward the following directions, which can provide directions for future AI research:

1) Machine learning (ML) technology and RNN-based algorithms are used to develop intelligent systems in large enterprise environments. Using different RNN structures, the optimal performance of the structure is obtained in threshold test to develop a suitable intelligent system.

2) Through the family of machine learning models, an optimal machine learning model is established. By choosing an appropriate machine learning model, the stage process of artificial intelligence application development can be displayed, and a set of practical applications or platforms based on the model can be further constructed.

3) Use the technology based on machine learning model to build a large-scale product recommendation system. Rigorous model construction makes the experimental project more accurate.

4) Session-based fraud detection in online e-commerce transactions can also be run by establishing a machine learning model. On this basis, the detection results of the information in the conversation can be obtained in the monitoring and evaluation stage.

5) In online e-commerce transactions, conversation-based fraud detection can adopt qualitative research, choose the best system architecture, and establish a conversation-based fraud detection system. It is more convenient for different professionals to use and can achieve popular fraud detection.

6) For the development of a set of artificial intelligence applications, RNN-based algorithms can be used. By comparing the cost and performance of the algorithm, the most suitable workflow is selected.

6 CONCLUSIONS

Our survey showed that the choice of computer architecture or platform is a key factor of AI deployment. That is, they are prerequisites for successful AI deployment. At the same time, there are many research challenges left due to the lack of quantitative data. In the future, optimal solutions for the deployment of artificial intelligence system should be developed.

REFERENCES

[1] Amerce, S., Bagel, A., Bird, C., Decline, R., Gall, H., Kamar, E., Nagappan, N., Noshi, B. and Zimmermann, T., 2019, May. Software engineering for machine learning: A case study. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 291-300). IEEE.

[2] Bernarda, L., Mavroudis, T. and Estevez, P., 2019, July. 150 successful machine learning models: 6 lessons learned at booking. com. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1743-1751).

[3] Kangal, B. and Tata, S., 2018, April. Recommendations for all: Solving thousands of recommendation problems daily. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 1404-1413). IEEE.

[4] Hill, C., Bellaour, R., Erickson, T. and Burnett, M., 2016, September. Trials and tribulations of developers of intelligent systems: A field study. In 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC) (pp. 162-170). IEEE.

[5] Wang, S., Liu, C., Gao, X., Qu, H. and Xu, W., 2017, September. Session-based fraud detection in online e-commerce transactions using recurrent neural networks. In Joint European Conference on

Machine Learning and Knowledge Discovery in Databases (pp. 241-252). Springer, Cham.

[6] Meena Mary John，Helena Holmstrom Olsson，Jan Bosch 2020. AI Deployment Architecture: Multi-Case Study for Key Factor Identification

[7] Lwakatare, L.E., Raj, A., Crnkovic, I., Bosch, J. and Olsson, H.H., 2020.Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. Information and Software Technology, 127, p.106368.

[8] John, M.M., Olsson, H.H. and Bosch, J., 2020. August. AI on the Edge: Architectural Alternatives. In 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (Accepted)

[9] Lucci, S. and Kopec, D., 2015. Artificial intelligence in the 21st century. Stylus Publishing, LLC.

[10] Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

[11] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., 2016. Deep learning (Vol. 1, p. 2). Cambridge: MIT press.

[12] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. and Dennison, D., 2015. Hidden technical debt in machine learning systems. In Advances in neural information processing systems (pp. 2503-2511).

[13] Dahlmeier, D., 2017, July. On the Challenges of Translating NLP Research into Commercial Products. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 92-96).

[14] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. and Dennison, D., 2015. Hidden technical debt in machine learning systems. In Advances in neural information processing systems (pp. 2503-2511).

[15] Maxwell, J.A., 2012. Qualitative research design: An interactive approach (Vol. 41). Sage publications.

[16] Walsham, G., 1995. Interpretive case studies in IS research: nature and method. European Journal of information systems, 4(2), pp.74-81.

[17] Runeson, P. and Host, M., 2009. Guidelines for conducting and reporting ¨ case study research in software engineering. Empirical software engineering, 14(2), p.131.

[18] Wilson, V., 2014. Research methods: triangulation. Evidence based library and information practice, 9(1), pp.74-75.

[19] Easterbrook, S., Singer, J., Storey, M.A. and Damian, D., 2008. Selecting empirical methods for software engineering research. In Guide to advanced empirical software engineering (pp. 285-311). Springer, London.