Learning Parameter Analysis for Machine Reading Comprehension

Xuekui Li, Lei Chen(✉), Yi Shi, and Ping Cui

Jiangsu Province Key Laboratory of Intelligent Industry Control Technology, Xuzhou University of Technology, Xuzhou 221018, China
chenlei@xzit.edu.cn

Abstract. Machine reading comprehension is a classic issue artificial intelligence. It is a key technology in the next generation search engine and intelligent interactive service. The traditional methods usually work in a small scale of data sets. The traditional system cannot meet the emerging demand. Deep learning and cloud computation have ability to deal with the large scale data sets. In real scene, the parameters affect the performance of machine reading comprehension task. In this paper, we analyze how the parameters of deep neural network affect the machine reading comprehension. The experiment results show that the performance is only sensitive to a few parameters which should be key point for engineers.

Keywords: Machine reading comprehension · BiDAF · Bleu · Rouge-I · Parameter analysis · Deep learning

1 Introduction

The rapid development of 5G and artificial intelligent technologies lead to the possibility of complex human-machine interaction in mobile real-time environment. The current communication researches focus on improving some key performance, such as energy-efficiency [1–3], spectral-efficiency [4–7], traffic flow optimization [8–11], and QoS guarantee [12–15]. To meet the requirements of different services in a large number of fields, some special optimization strategies are proposed [16–18]. The new strategies can effectively improve the user experience [19–22] in the basis of user behavior analysis [23, 24] and traffic prediction [25, 26]. However, for some special human-machine communication services, the user experience is not like the traditional service that is affected by only the effective capacity of links [27–30]. The cloud computation capacity, data set quality, and interactive algorithm can also affect the end user experience [31]. With limited computation capacity, the algorithm efficiency is important for user experience improvement. And a lot of parameters might cause performance fluctuation [32–34].

In this paper, we investigate how the deep learning parameters affect the performance on DUREADER that is an open-access Chinese machine reading comprehension data

H. Song and D. Jiang (Eds.): SIMUtools 2020, LNICST 369, pp. 485–494, 2021.
https://doi.org/10.1007/978-3-030-72792-5_39

set. Compared with the previous MRC data set, DUREADER has the following characteristics. 1) All questions and original texts are actual data collected from the Baidu search engine data and Baidu community; 2) The data set contains both a large number of right and wrong samples that were rarely studied before; 3) Each question corresponds to multiple answers. As the largest Chinese MRC data at present, DUREADER set contains 200K questions, 1000K original texts and 420k answers.

2 BiDAF Model

In the process of machine reading comprehension, we will give a question *(q)* and one or more paragraphs *(P)* / document *(d)*, and then use the machine to find the correct answer *(a)* in the given paragraphs, that is, $q + P \; or \; D => A$. machine reading comprehension is one of the key tasks in natural language processing (NLP), which requires the machine to have a deep understanding of the language to find the correct answer. We adopt paddle-fluid as tools to implement the classic reading comprehension model – BiDAF [35] on DUREADER [36].

The model diagram is shown in Fig. 1.

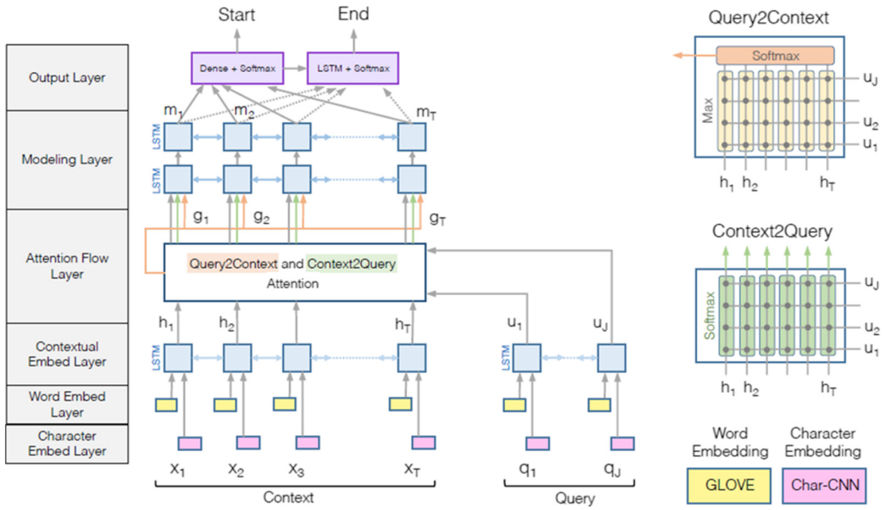


Fig. 1. The model diagram [35].

The model is a multi-layer process with six layers of network [35].

(1) Character embed layer: Mapping each word to vector space with character level CNN.
(2) Word embed layer: Using the pre-trained word embedding model, each word is mapped to a vector space. After embedding and splicing characters and words, input them to a double-layer highway network, and the output is two d-Dimension vectors [37]. Mark the output of the original as X $\in R^{d*t}$, and the output of the question as Q $\in R^{d*t}$ [35].

(3) Context embed layer: Use context clues of surrounding words to refine the embedding of words. The first three layers are applied to both the question and the original.
(4) Attention flow layer: The question vector is coupled with the original vector, and a set of feature vectors related to the question is generated for each word in the original. The above three layers are mainly used to extract information of different levels and granularity from text and query, and this layer is used to link and integrate text and query information [38].
(5) Modeling layer: Use RNN to scan the entire text. The input of this layer of network is the output G of attention flow layer (a text word representation of query aware). The whole network can capture the relationship between text words based on the query. This is different from the contextual embedding layer, which is not based on queries to capture the relationship between text words.
(6) Output layer: Output answers corresponding to questions. The output layer is a specific network structure. The BIDAF network allows you to modify the output layer according to specific tasks, but keep other layer structures. So we set up a specific output layer for QA tasks.

3 Experiment and Analysis

Through the experimental study of machine reading, the parameters (learning rate, weight decay, hidden size, embedded size) in the model are tested to see how the changes of parameters affect the running results and performance (bleu-4, rouge-l) [39, 40]. First, the following parameters are introduced:

(1) Learning rate: the size of learning rate. The type is float. The default value is 0.001.
(2) Weight decay: weight attenuation, type is float, default is 0.0001.
(3) Hidden_size: the size of running hidden units. The data type is int, and the default is 300.
(4) Embed_size: the dimension of the embedded table. The data type is int, and the default value is 300.

The following is the calculation formula and meaning of the result:

The full name of Bleu is bilingual evaluation understanding. The score range of Bleu is 0–1. The closer the score is to 1, the higher the quality of translation. Bleu is mainly based on precision. The following is the overall formula of Bleu [39]:

$$\begin{aligned} \mathrm{BLEU} &= BP \times \exp\left(\sum_{n=1}^{N} W_n \times \log P_n\right) \\ BP &= \begin{cases} 1, & lc > lr \\ exp\left(1 - lr/lc\right), & lc \le lr \end{cases} \end{aligned} \tag{1}$$

Bleu needs to calculate the accuracy of translation 1-g, 2-g, …, n-gram. PN in the formula refers to the accuracy of n-gram. Wn refers to the weight of n-gram, which is generally set as uniform weight, that is, for any n, WN = 1/n [4]. BP is the penalty factor, if the length of the translation is less than the shortest reference translation, then BP is less than 1. LC is the length of machine translation, LR is the shortest length of reference translation sentence. Bleu's 1-g accuracy indicates the degree of faithfulness of the translation to the original, while other n-grams indicate the fluency of the translation.

The full name of the rouge indicator is "recall oriented under study for giving evaluation", which is mainly based on the recall rate. Rouge is a commonly used evaluation index of machine translation and abstracts of articles, which is proposed by Chin yew Lin. Four Rouge methods are proposed in this paper [40]: 1) Rouge-n: calculate the recall rate on n-gram. 2) Rouge-l: the longest common subsequence between machine translation and reference translation is considered. 3) Rouge-w: improved rouge-l to calculate the longest common subsequence by weighted method.

In this paper, we employ the rouge-l method. L in rouge-l refers to the longest common subsequence (LCS). The calculation formula is as follows:

$$\begin{aligned} R_{LCS} &= \frac{LCS(C,S)}{len(S)} \\ P_{LCS} &= \frac{LCS(C,S)}{len(C)} \\ F_{LCS} &= \frac{\left(1+\beta^2\right)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \end{aligned} \tag{2}$$

RLCs in the formula represents the recall rate, while PLCs represents the accuracy rate, and FLCs is the rouge-l. Generally, beta is set to a large number, so FLCs only considers RLCs (recall rate) [5]. Note that if the beta is large, then f will pay more attention to r than P. see the formula below. If the beta is large, then PLCs is negligible [40].

$$\frac{1}{F_{LCS}} = \frac{1}{\left(1+\beta^2\right)P_{LCS}} + \frac{\beta^2}{(1+\beta 2)R_{LCS}} \tag{3}$$

Next is the test chart of the experimental parameters:

The experimental parameters are listed in Table 1, 2, 3 and Table 4.

The test result are shown in Fig. 2.

It can be seen from the figure that different parameter changes cause different levels of fluctuation. It shows that the changes of learning_rate and Weight_decay cause a non-monotonous fluctuation. Engineers should pay more attention to adjust these parameters. And some parameters, such as hidean_size, does not cause acute performance fluctuation. The developer noly need to avoid the worst point.

Table 1. Learning_rate experimental parameters.

Learning_rate	Bleu-4	Rouge-L
0.00022	0.301159816	0.335547702
0.000222	0.327446346	0.349522438
0.000223	0.310633997	0.343745255
0.000224	0.309503916	0.345074965
0.00022425	0.314377135	0.3490552
0.0002245	0.300811443	0.341126438
0.000225	0.299887976	0.346905289
0.00023	0.275785047	0.325982007
0.00024	0.305897563	0.338380032
0.00025	0.260849998	0.326924794

Table 2. Weight_decay experimental parameters.

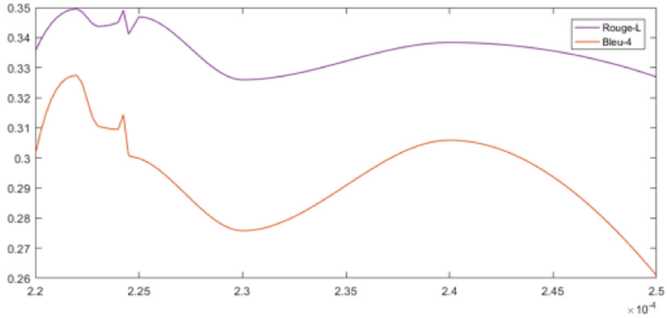
Weightdecay	Bleu-4	Rouge-L
0.0002	0.268626609	0.332181801
0.000225	0.254432416	0.332603854
0.00025	0.268450285	0.336462598
0.000275	0.301540809	0.355285356
0.0003	0.273552474	0.340523308
0.000325	0.253350986	0.333022634
0.00035	0.280247462	0.334507126
0.000375	0.272113977	0.337251798
0.0004	0.301491039	0.344305035
0.000425	0.318691688	0.355967577
0.00045	0.246738094	0.329907082
0.000475	0.285758484	0.346616667
0.0005	0.268979783	0.336884469

Table 3. Hidden_size experimental parameters.

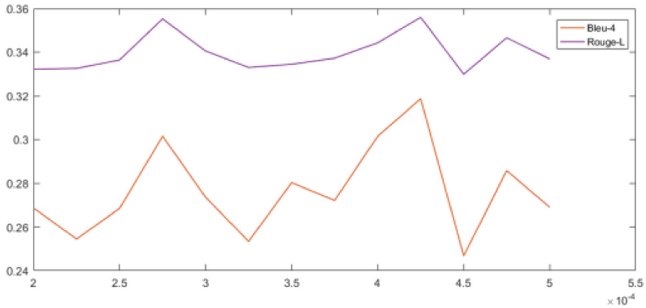
Hiddensize	Bleu-4	Rouge-L
300	0.287737556	0.342004994
350	0.297440165	0.347887622
400	0.174288674	0.30164207
450	0.303693193	0.35553517
500	0.273745151	0.339456535
550	0.293851017	0.341371026
600	0.317234853	0.353995966
650	0.318028688	0.3578506
700	0.303053313	0.34409938
750	0.290282375	0.34627556

Table 4. Embed_size experimental parameters.

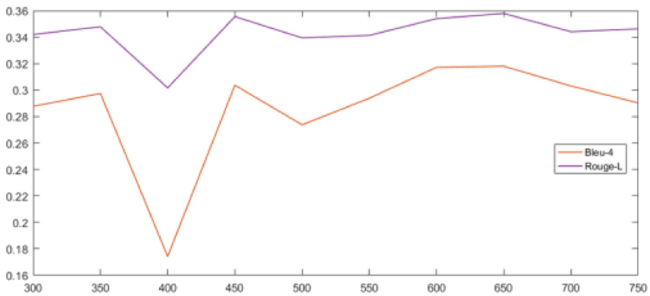
Embedsize	Bleu-4	Rouge-L
300	0.274609626	0.343514817
350	0.275633768	0.342759499
400	0.268605629	0.338488701
450	0.27385542	0.336640192
500	0.308942148	0.351445457
550	0.247504463	0.332083785
600	0.250511795	0.320621626
650	0.276932374	0.339581689
700	0.309809729	0.349099037
750	0.298543118	0.346263549



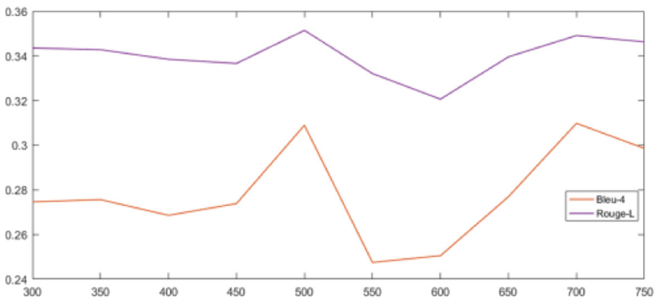
(a) Learning_rate



(b) Weight_decay



(c) Hidden_size



(d) Embed_size

Fig. 2. Parameters and performances.

4 Conclusion

Through the experiments, it is found that the parameters used in machine reading affect the accuracy and time efficiency in different way. For each parameter, we need to have different adjust strategy. Some parameters cause periodical fluctuation, we only need to adjust in a small interval. For the parameter that causes a flat performance change, we only need to find the worst point and avoid it. The main work is to adjust the parameters that cause irregular performance fluctuation.

Acknowledgements. This work is partly supported by Jiangsu major natural science research project of College and University (No. 19KJA470002).

References

1. Jiang, D., Wang, Y., Lv, Z., Wang, W., Wang, H.: An energy-efficient networking approach in cloud services for IIoT networks. IEEE J. Sel. Areas Commun. **38**(5), 928–941 (2020)
2. Jiang, D., Zhang, P., Lv, Z., et al.: Energy-efficient multi-constraint routing algorithm with load balancing for smart city applications. IEEE Internet Things J. **3**(6), 1437–1447 (2016)
3. Jiang, D., Li, W., Lv, H.: An energy-efficient cooperative multicast routing in multi-hop wireless networks for smart medical applications. Neurocomputing **220**(2017), 160–169 (2017)
4. Lee, Y., Kim, Y., Park, S.: A machine learning approach that meets axiomatic properties in probabilistic analysis of LTE spectral efficiency. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea (South), pp. 1451–1453 (2019)
5. Ji, H., Sun, C., Shieh, W.: Spectral efficiency comparison between analog and digital RoF for mobile fronthaul transmission link. J. Lightwave Technol. **38**, 5617–5623 (2020)
6. Hayati, M., Kalbkhani, H., Shayesteh, M.G.: Relay selection for spectral-efficient network-coded multi-source D2D communications. In: 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, pp. 1377–1381 (2019)
7. You, L., Xiong, J., Zappone, A., Wang, W., Gao, X.: Spectral efficiency and energy efficiency tradeoff in massive MIMO downlink transmission with statistical CSIT. IEEE Trans. Signal Process. **68**, 2645–2659 (2020)
8. Jiang, D., Wang, W., Shi, L., Song, H.: A compressive sensing-based approach to end-to-end network traffic reconstruction. IEEE Trans. Netw. Sci. Eng. **7**(1), 507–519 (2020)
9. Wang, Y., Jiang, D., Huo, L., Zhao, Y.: A new traffic prediction algorithm to software defined networking. Mob. Netw. Appl. (2019)
10. Qi, S., Jiang, D., Huo, L.: A prediction approach to end-to-end traffic in space information networks. Mob. Netw. Appl. 1–10 (2019)
11. Huo, L., Jiang, D., Lv, Z., et al.: An intelligent optimization-based traffic information acquirement approach to software-defined networking. Comput. Intell. **36**, 1–21 (2019)
12. Tan, J., Xiao, S., Han, S., Liang, Y., Leung, V.C.M.: QoS-aware user association and resource allocation in LAA-LTE/WiFi coexistence systems. IEEE Trans. Wireless Commun. **18**(4), 2415–2430 (2019)
13. Wang, Y., Tang, X., Wang, T.: A unified QoS and security provisioning framework for wiretap cognitive radio networks: a statistical queueing analysis approach. IEEE Trans. Wireless Commun. **18**(3), 1548–1565 (2019)

14. Hassan, M.Z., Hossain, M.J., Cheng, J., Leung, V.C.M.: Hybrid RF/FSO backhaul networks with statistical-QoS-aware buffer-aided relaying. IEEE Trans. Wireless Commun. **19**(3), 1464–1483 (2020)
15. Zhang, Z., Wang, R., Yu, F.R., Fu, F., Yan, Q.: QoS aware transcoding for live streaming in edge-clouds aided HetNets: an enhanced actor-critic approach. IEEE Trans. Veh. Technol. **68**(11), 11295–11308 (2019)
16. Jiang, D., Huo, L., Lv, Z., Song, H., Qin, W.: A joint multi-criteria utility-based network selection approach for vehicle-to-infrastructure networking. IEEE Trans. Intell. Transp. Syst. **19**(10), 3305–3319 (2018)
17. Jiang, D., Wang, Y., Lv, Z., Qi, S., Singh, S.: Big data analysis based network behavior insight of cellular networks for industry 4.0 applications. IEEE Trans. Industr. Inf. **16**(2), 1310–1320 (2020)
18. Wang, F., Jiang, D., Qi, S.: An adaptive routing algorithm for integrated information networks. China Commun. **7**(1), 196–207 (2019)
19. Barakabitze, A.A., et al.: QoE management of multimedia streaming services in future networks: a tutorial and survey. IEEE Commun. Surv. Tutor. **22**(1), 526–565 (2020)
20. Orsolic, I., Skorin-Kapov, L.: A framework for in-network QoE monitoring of encrypted video streaming. IEEE Access **8**, 74691–74706 (2020)
21. Song, E., et al.: Threshold-oblivious on-line web QoE assessment using neural network-based regression model. IET Commun. **14**(12), 2018–2026 (2020)
22. Seufert, M., Wassermann, S., Casas, P.: Considering user behavior in the quality of experience cycle: towards proactive QoE-aware traffic management. IEEE Commun. Lett. **23**(7), 1145–1148 (2019)
23. Jiang, D., Huo, L., Song, H.: Rethinking behaviors and activities of base stations in mobile cellular networks based on big data analysis. IEEE Trans. Netw. Sci. Eng. **7**(1), 80–90 (2020)
24. Bao, R., Chen, L., Cui, P.: User behavior and user experience analysis for social network services. Wireless Netw. (2020). https://doi.org/10.1007/s11276-019-02233-x
25. Jiang, D., Huo, L., Li, Y.: Fine-granularity inference and estimations to network traffic for SDN. PLoS ONE **13**(5), 1–23 (2018)
26. Huo, L., Jiang, D., Qi, S., et al.: An AI-based adaptive cognitive modeling and measurement method of network traffic for EIS. Mob. Netw. Appl. (2019).
27. Guo, C., Liang, L., Li, G.Y.: Resource allocation for low-latency vehicular communications: an effective capacity perspective. IEEE J. Sel. Areas Commun. **37**(4), 905–917 (2019)
28. Shehab, M., Alves, H., Latva-aho, M.: Effective capacity and power allocation for machine-type communication. IEEE Trans. Veh. Technol. **68**(4), 4098–4102 (2019)
29. Cui, Q., Gu, Y., Ni, W., Liu, R.P.: Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: from theory to application. IEEE J. Sel. Areas Commun. **35**(8), 1754–1767 (2017)
30. Xiao, C., Zeng, J., Ni, W., Liu, R.P., Su, X., Wang, J.: Delay guarantee and effective capacity of downlink NOMA fading channels. IEEE J. Sel. Top. Sign. Process. **13**(3), 508–523 (2019)
31. Zhang, K., Chen, L., An, Y. et al.: A QoE test system for vehicular voice cloud services. Mobile Netw. Appl. (2019). https://doi.org/10.1007/s11036-019-01415-3
32. Farooq, H., Kaushik, B.: Review of deep learning techniques for improving the performance of machine reading comprehension problem. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 928–935 (2020)
33. Guo, J., Liu, G., Xiong, C.: Multiple attention networks with temporal convolution for machine reading comprehension. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, pp. 546–549 (2019)

34. Jin, W., Yang, G., Zhu, H.: An efficient machine reading comprehension method based on attention mechanism. In: 2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, pp. 1297–1302 (2019)
35. Seo, M., Kembhavi. A., Farhadi. A., et al.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
36. He, W., Liu. K., Liu, J., et al.: DuReader: a Chinese machine reading comprehension dataset from real-world applications. In: Association for Computational Linguistics, Special Issue: Proceedings of the Workshop on Machine Reading for Question Answering, pp. 37–46 (2018)
37. Etzioni, O., Banko, M., Cafarella, M.J.: Machine reading. In: AAAI Spring Symposium: Machine Reading, Technical Report, Stanford, California, USA, DBLP, pp. 1–5 (2007)
38. Shen, Y., Huang, P.S., Gao, J., et al. ReasoNet: learning to stop reading in machine comprehension. In: Proceedings of the 23rd ACM SIGKDD International Conference. ACM (2017)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics(ACL), Philadelphia, pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135
40. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, pp. 74–81 (2004). https://www.aclweb.org/anthology/W04-1013