Multi-task based Neural Network for Cloud Service Recommendation

Ousainou Jaiteh
Computer Science Department
University of The Gambia (UTG)

Abstract

This research proposes a Multi-task based Neural Network for Cloud Service Recommendation (M-Rec) approach. M-Rec utilizes the user's historical interactions with cloud services to provide personalized recommendations by jointly predicting the response time and throughput service category. The model is evaluated on the WS-DREAM dataset and compared to state-of-the-art recommendation methods in terms of Root Mean Square Error (RMSE)and Mean Absolute Error (MAE). The experimental results demonstrate that M-Rec outperforms most baseline methods and is also shown to be scalable and efficient.

Introduction

- The study of recommendation and selection for cloud services over the years. Several related works have been proposed and implemented to improve recommendation techniques and expand application scenarios, resulting in remarkable recommendation performance.
- To understand the development process and aid further model improvement, an empirical study was conducted to investigate influential works in this field, including Collaborative filtering, Factorization Machine, Matrix factorization, Machine Learning, Deep Learning, Natural Language Processing, Optimization Algorithm, and Graph-based models.
- Additionally, a multi-task-based approach was employed to predict multiple tasks simultaneously, specifically response time and throughput.

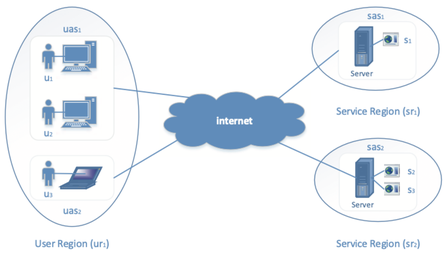

Figure 1. The interactions between end users and service systems where the round corner rectangle represents an autonomous system and the ellipse indicates a country/region.

Methodology

- The Mec-Rec relies on multiple existing techniques, which include word embedding, neural networks, and multi-task learning.
- The use of the context-sensitive text representation method word2vec [1][2][3] to vectorize the input and convert the text input value into a vector with continuous numeral values of input values.

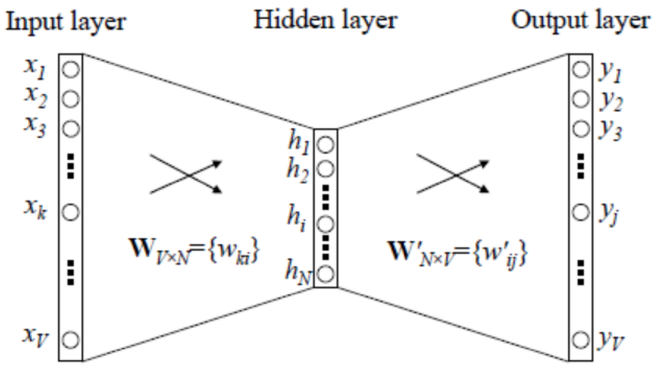

Figure 2. Structure of Word2Vec.

- The M-Rec model which is inspired by an approach of multi-task learning is centered on a new structure called the Multi-gate Mixture-of-Experts (MMOE)[4].
- This structure explicitly represents the connections between tasks and learns functions that are specific to each task, all while utilizing shared representations.
- By automatically allocating parameters, the structure can capture shared and task-specific information without requiring the addition of numerous new parameters for each task.
- Each expert network is a unique shared bottom network, each using the same architecture.

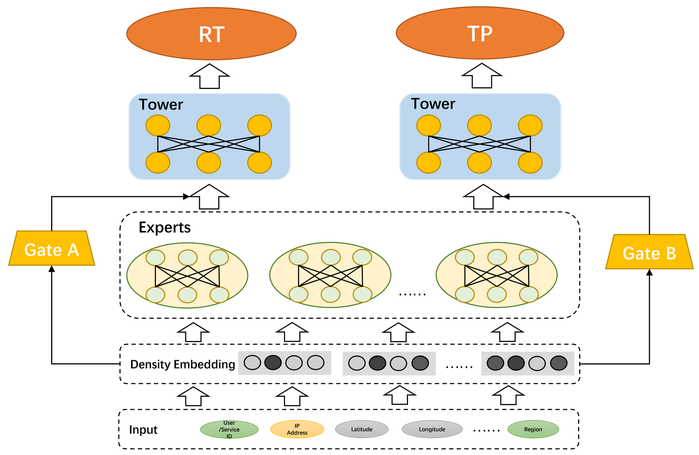

Figure 3. Structure of the M-Rec.

Dataset and Experiment Settings

- All the experiments are conducted on a high-performance computer with 2 CPUs (Intel Xeon Gold 6142), 4 RAMs (32G DDR4 2666MT/s) and 2 GPU cards (NVIDIA GeForce RTX 3070 Ti, 8G), running the Ubuntu 16.04 OS. The DSSN is implemented upon the popular deep learning framework Tensorflow 1.14.0 1.
- The attributes of users and services, such as IP, country, and description, serve as the original features of each graph entity, which are preprocessed to build the initial embedding vectors.
- A the public real-world dataset WS-Dream 2 as an experimental dataset, which contains 339 users, 5825 web services, 1, 873, 828 records for response time, and 1, 831, 252 records for throughput. The densities of the response time and throughput are 0.94893 and 0.927369 respectively.

Table 1. Outline of QoS Dataset.

Attribute	#Users	#Services	Range	Density
RT (sec)	339	5825	0-20	92.7%
TP (kbps)	339	5825	0-1000	94.9%

Results

To demonstrate the validity of the proposed M-Rec, we compare the experimental results of service recommendation with 6 typical baseline models, including UPCC [5], IPCC [6], UIPCC [6], PMF [7], FM [8].

- UPCC is a typical CF-based recommendation model, which recommends the objects using user-based similarity computing.
- IPCC is also a typical CF-based recommendation model, which recommends the objects using item-based similarity computing.
- UIPCC takes the advantages of both UPCC and IPCC, and combines the results of the two models to obtain a comprehensive prediction.
- PMF is a typical MF-based recommendation model for QoS-based prediction. Specifically, the PMF improves the traditional MF model with probabilistic settings.
- FM directly uses the factorization machine model to make the QoS prediction. Specifically, the FM exploits the categorical interactions with feature domains to obtain the embedding vectors.

Table 2. Comparison of QoS Prediction.

2*QoS Attributes	2*Methods	Density=2.5%		Density=5%		Density=10%	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
6*Response Time	UPCC	0.726	1.52	0.65	1.41	0.585	1.348
	IPCC	0.763	1.684	0.647	1.422	0.595	1.357
	UIPCC	0.717	1.51	0.637	1.412	0.583	1.347
	PMF	0.646	1.726	0.597	1.541	0.543	1.428
	FM	0.711	1.524	0.623	1.428	0.55	1.341
	M-Rec*	0.703	1.522	0.583	1.402	0.524	1.334
6*Throughput	UPCC	28.73	63.06	25.27	56.01	20.36	50.3
	IPCC	27.79	66.15	23.56	56.43	22.80	54.21
	UIPCC	28.14	62.45	24.83	55.94	20.12	50.18
	PMF	25.46	76.5	22.77	70.69	21.58	67.02
	FM	25.67	70.8	23.41	62.8	20.89	57.87
	M-Rec*	24.19	60.78	22.82	54.79	19.32	48.84

Conclusion

- In conclusion, this work implemented a multi-task based recommendation model for cloud service recommendation M-Rec. The M-Rec receives the input that is preprocessed by the typical word embedding technique and then feeds the vector input into several expert networks.
- It's trained by two network towers for different recommendation metrics which are throughput and response time. The experimental results show that our model outperforms traditional recommendation methods in terms of MAE and RMSE.

References

[1] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," arXiv preprint arXiv:1402.3722, 2014.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of Advances in Neural Information Processing Systems, pp. 3111–3119, 2013.

[4] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1930–1939, 2018.

[5] L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, and H. Mei, "Personalized qos prediction for web services via collaborative filtering," in Ieee international conference on web services (icws 2007), pp. 439–446, IEEE, 2007.

[6] Z. Zheng, H. Ma, M. R. Lyu, and I. King, "Qos-aware web service recommendation by collaborative filtering," IEEE Transactions on services computing, vol. 4, no. 2, pp. 140–152, 2010.

[7] Z. Zheng, H. Ma, M. R. Lyu, and I. King, "Collaborative web service qos prediction via neighborhood integrated matrix factorization," IEEE Transactions on Services Computing, vol. 6, no. 3, pp. 289–299, 2012.

[8] S. Rendle, "Factorization machines with libfm," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 3, pp. 1–22, 2012.

ojaiteh@utg.edu.gm