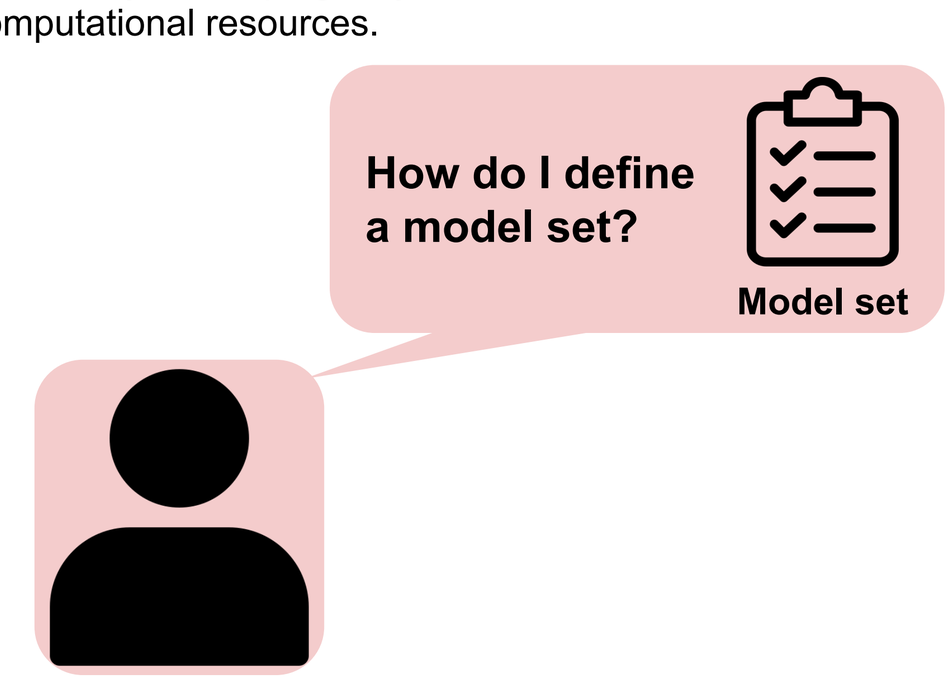
Toward Model Selection Through Measuring Dataset Similarity on TensorFlow Hub

SeungYoung Oh[1], Hyunmin Lee[1], JinHyun Han[1,2], Hyunggu Jung[1]

[1]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, UNIVERSITY OF SEOUL; [2]UST21

INTRODUCTION

- The purpose of this study is to develop a model selection system that addresses the increasing computational resource consumption in the machine learning model development process.
- Current model selection systems primarily focus on selecting models from a given set without addressing the method of determining the model set.
- Therefore, our goal is to assist developers in the model selection process by developing a system that reduces the use of computational resources.



METHODS

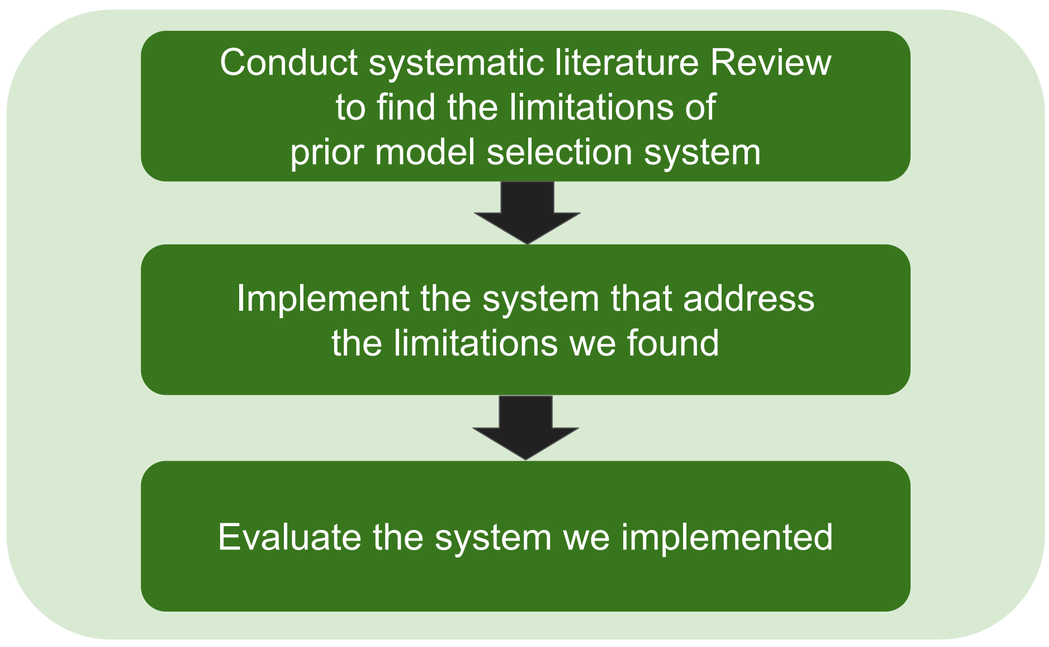


Fig 1. The flow chart of this study.

- The research objective is to develop and evaluate a model selection system.
- First, we conducted a literature review to identify the limitations of existing model selection systems and compiled a list of features to overcome these limitations.
- Next, we developed the model selection system using the data from TF Hub and implemented the features.
- Finally, we evaluated the implemented system to determine how many limitations were overcome based on the limitations identified through the literature review.

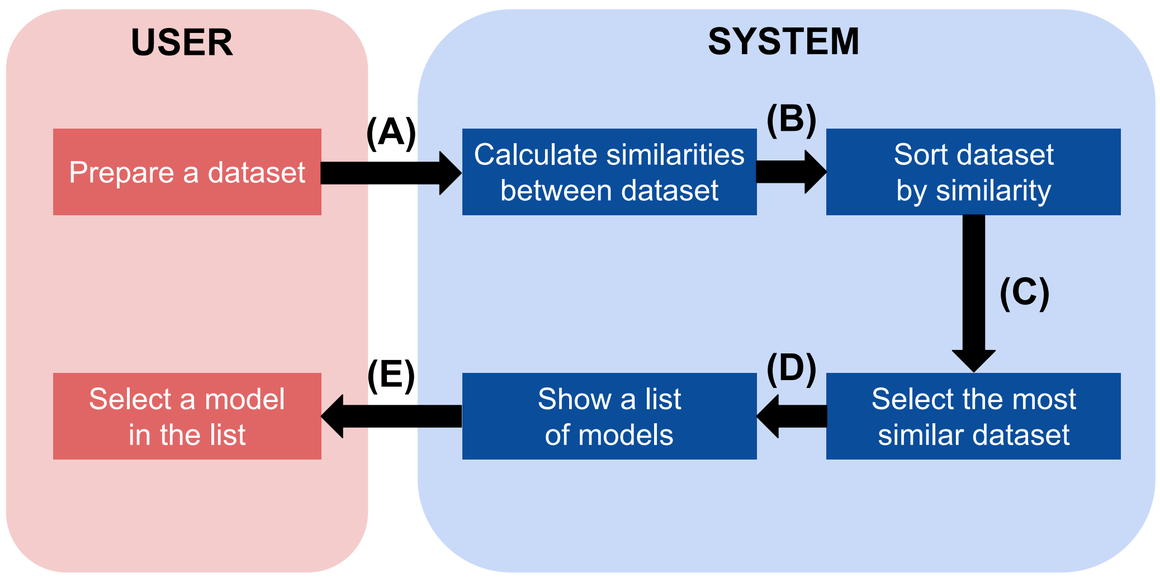


Fig 2. The flowchart of our model selection system. The user enters the dataset (A) that the user wants to train. The system calculates the similarity between the datasets and generates the similarity list (B). After sorting the list (C) by the similarity value, the most similar dataset name (D) is determined. Finally, the user selects the model from the list (E) which contains the models previously trained with (D).

Further information

Author Homepage QR Code

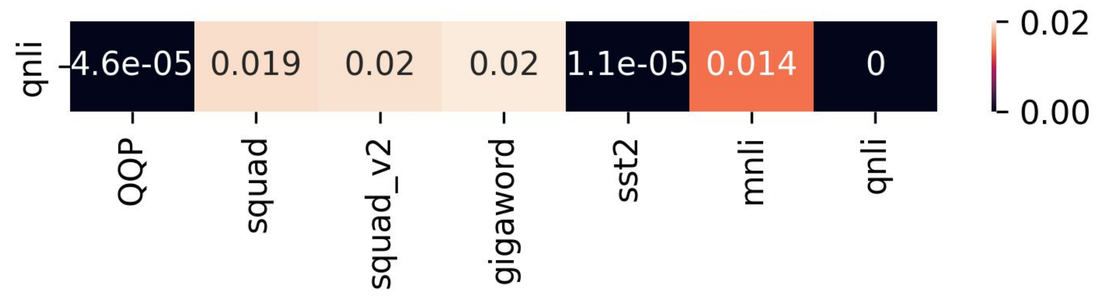


Fig 3. The heat map of similarity between datasets. **(B)** of **Fig 2**. The similarity between datasets was measured using the cosine similarity formula.

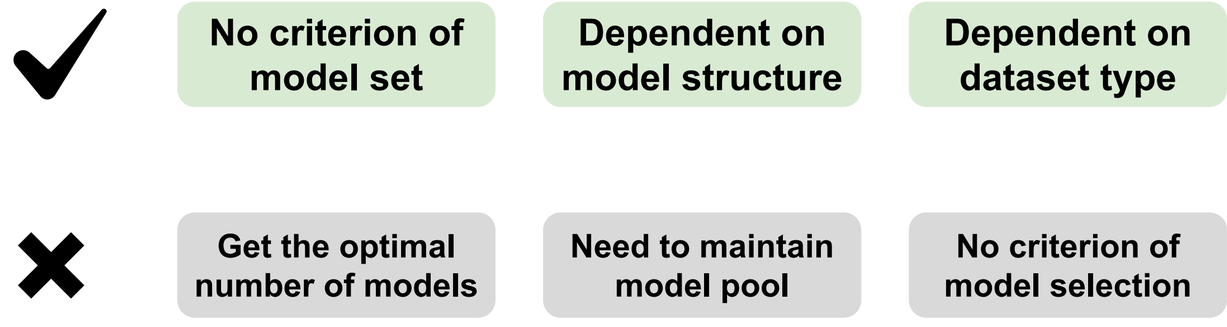


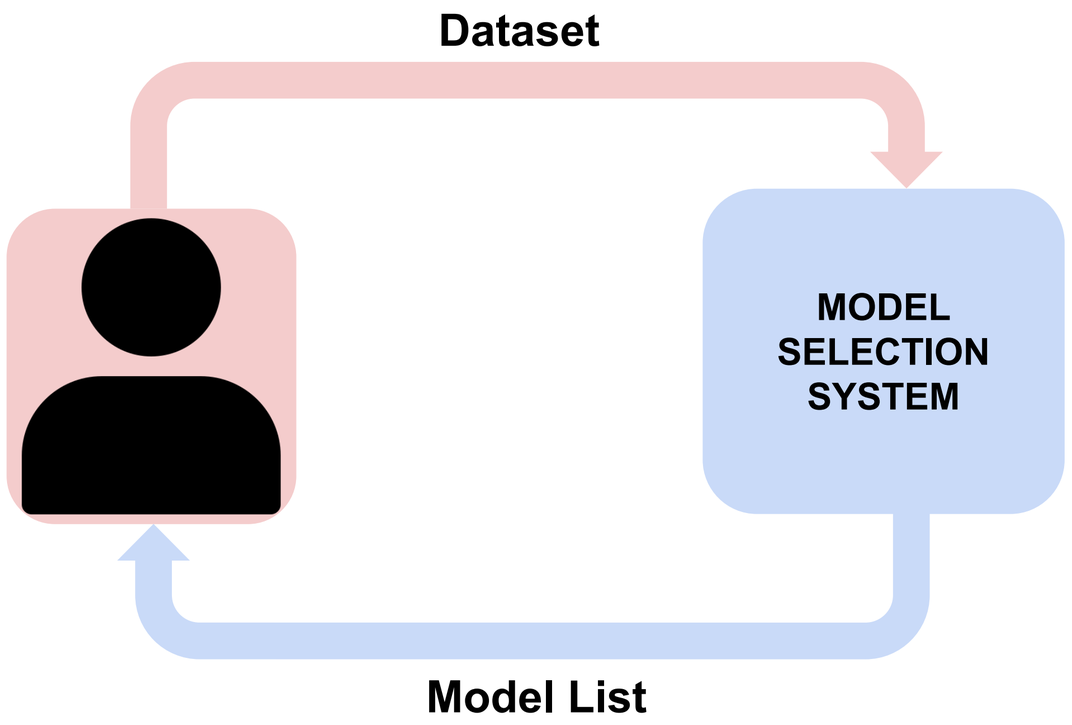
Fig 4. The limitations of prior model selection studies. Our system addressed 3 of the 6 identified limitations. Green squares are the addressed limitations and gray squares are the unaddressed limitations.

RESULTS

- Our proposed system addresses three of the six identified limitations.
- To evaluate how many limitations our system has overcome, we create a spreadsheet with limitations in previous studies (Figure 4).
- Our proposed system successfully addressed three limitations through the following features:
  1. Automatically collects models from the TF Hub
  2. Does not depend on the model structure, as it recommends models based on the similarity of the dataset and the datasets on which the model was trained.
  3. Accepts any type of data if it is transformed into a tabular format of numbers.

CONCLUSION

- The ultimate goal of this study is to develop a model selection system that suggests an AI model without any prior background knowledge.
- To achieve our goal, we crawled the information about AI models and datasets from TF Hub.
- Using the information, we measured the similarity between the datasets.
- By using the similarity, our system suggests an AI model.
- In this paper, we propose a system that automatically retrieves model sets and recommends artificial intelligence models.
- The developed system assists developers in selecting models, thereby reducing the consumption of computational resources and challenges in model selection.



REFERENCES

1. Abdallah, M., Rossi, R., Mahadik, K., Kim, S., Zhao, H., & Bagchi, S. (2022, October). AutoForecast: Automatic Time-Series Forecasting Model Selection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 5-14).
2. Jomaa, H. S., Schmidt-Thieme, L., & Grabocka, J. (2021). Dataset2vec: Learning dataset meta-features. *Data Mining and Knowledge Discovery*, *35*, 964-985.
3. Yoo, T., Chun, M., Bae, Y., Kwon, S., & Jung, H. (2022, November). Exploring the Community of Model Publishers on TensorFlow Hub. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 63-67).

CONTACT SeungYoung Oh (dhtmddud08@gmail.com); Hyunggu Jung (hjung@uos.ac.kr)