HAPI Explorer: Comprehension, Discovery, and Explanation on History of ML APIs

Lingjiao Chen[1*], **Zhihua Jin**[2*], **Sabri Eyuboglu**[1], **Huamin Qu**[2], **Christopher Ré**[1], **Matei Zaharia**[1], **James Zou**[1]

[1] Stanford University
[2] Hong Kong University of Science and Technology
lingjiao@stanford.edu, zjinak@cse.ust.hk, Eyuboglu@stanford.edu, huamin@cse.ust.hk, chrismre@cs.stanford.edu,
matei@cs.stanford.edu, jamesz@stanford.edu

Abstract

Machine learning prediction APIs offered by Google, Microsoft, Amazon, and many other providers have been continuously adopted in a plethora of applications, such as visual object detection, natural language comprehension, and speech recognition. Despite the importance of a systematic study and comparison of different APIs over time, this topic is currently under-explored because of the lack of data and user-friendly exploration tools. To address this issue, we present HAPI Explorer (History of API Explorer), an interactive system that offers easy access to millions of instances of commercial API applications collected in three years, prioritizes attention on user-defined instance regimes, and explain interesting patterns across different APIs, subpopulations, and time periods via visual and natural languages. HAPI Explorer can facilitate further comprehension and exploitation of ML prediction APIs.

Introduction

Machine learning (ML) prediction APIs offered by Google, Microsoft, Amazon, and many other providers have been continuously adopted in a plethora of applications, such as visual object detection, natural language comprehension, and speech recognition (Chen, Zaharia, and Zou 2020; Tramèr et al. 2016). Reliable ML deployments require systematical understanding and comparison of different APIs in varying aspects, including overall accuracy (Koenecke et al. 2020), inherent biases (Buolamwini and Gebru 2018), and prediction consistency over time (Chen, Zaharia, and Zou 2021). Besides global performance metrics, it is also increasingly critical to reason and explain ML APIs' performance on certain subpopulations. However, this topic is currently under-explored due to a lack of data and exploration tools.

To address this issue, we present HAPI Explorer (History of API Explorer), a web UI-based interactive system for programming-free comprehension, discovery, and explanation of ML APIs. Our design goal includes (i) diverse coverage of ML APIs and datasets over time, (ii) convenient evaluation metric customization, and (iii) flexible user-define data regime visualization.

For diverse coverage, HAPI Explorer is built on top of HAPI (Chen et al. 2022), a longitudinal database consisting of more than 1 million instances of commercial API applications collected over three years. This covers ML APIs provided by Google, Microsoft, Amazon, IBM, and other companies for a range of tasks. HAPI Explorer also allows users to upload their own collections of API instances. Standard accuracy, confidence shift, consistency over time, and group disparity are provided as the default metric to visualize, and HAPI Explorer also supports users to create their own metrics. For flexible attention prioritization, HAPI Explorer enables "zoom in/zoom out" analysis: users can overlook the global performance view (Figure 1(a)), query a particular data regime to visualize (Figure 1(b)), or stare at ML APIs' predictions on a single data instance (Figure 1(c)). HAPI Explorer is available at https://hapi-explore.github.io/. We hope HAPI Explorer can stimulate more research and exploitation of ML APIs.

Related Work

The dynamics of API changes have been studied for a while as the MLaaS market grows bigger (Chen, Zaharia, and Zou 2020). They mainly focus on the prices of calling APIs (Chen, Koutris, and Kumar 2019) and prediction results analysis of the APIs (Hosseini, Xiao, and Poovendran 2017). As the changes in APIs happen, monitoring the ML pipeline is also important. However, they do not focus on longitudinal analysis of APIs, as the APIs will evolve over time, but they do not consider this aspect currently.

Research on MLaaS analysis focuses on the adaptive calling strategy for APIs and dataset evaluation for ML APIs. FrugalML (Chen, Zaharia, and Zou 2020) and FrugalMCT (Chen, Zaharia, and Zou 2022) are two kinds of adaptive strategies for calling the APIs, which enhance efficiency and lower the costs of using APIs. HAPI (Chen et al. 2022) is a recently proposed dataset for large-scale evaluation of ML APIs over time. Our demonstration is based on the HAPI dataset to enable users to understand the dynamics of API changes, which subsequently adapt their ML pipelines.

There are also a lot of visual interfaces developed for understanding and analyzing the models. The work on them is mainly categorized into three classes, model understanding, model diagnosis, and model refinement (Choo and Liu

[*] Co-first authors.

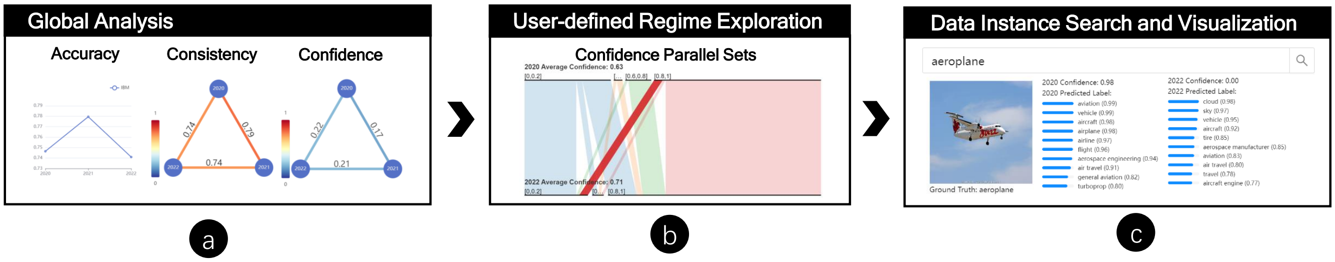

Figure 1: Main features of HAPI Explorer. (a) Global performance analysis helps visualize the metrics in the dataset and APIs. (b) Users can customize their choices on the confidence ranges in each year. (c) Users can search instances based on provided ground truth label. Corresponding instances are displayed.

2018). Since our work targets general model users, our work falls into the work of model understanding. This kind of work can be further categorized into model-agnostic and model-specific work. Model-agnostic work means that the visualization system is designed for analyzing general models instead of targeting one kind of model. For example, Squares (Ren et al. 2016) supports analyzing multi-class classification algorithm performance via class and instance-level distribution plots and bi-directional linking between the visualization and the corresponding instances. RegressionExplorer (Dingen et al. 2018) is an interactive visualization tool for logistic regression models through subgroup analysis of model performance. However, they do not focus on visualizing and understanding the dynamics of API changes. Such work cannot be directly applied to understand the dynamics of API changes.

Main Features of HAPI Explorer

To ease the study and research on ML prediction APIs, HAPI Explorer provides hierarchical access to a large scale of commercial prediction API application instances. Specifically, its main features are (i) global performance analysis (including accuracy, fairness and prediction consistency) for each evaluated dataset and ML API (Figure 1(a)), (ii) user-defined subpopulation (such as particular data class and confidence score range) exploration (Figure 1(b)), and (iii) efficient individual data instance (e.g., a single image) search on the entire database (Figure 1(c)). We discuss each feature in detail as follows.

Global performance analysis. Global performance analysis visually compares different ML APIs' overall performance over time across different datasets. Four different metrics are considered, including standard accuracy, prediction consistency, confidence shifts, and group disparity. Given the task, dataset and ML API specified by a user, HAPI Explorer produces different visual views of those metrics. Specifically, the accuracy plot is a line chart where the x-axis encodes the year and the y-axis encodes the performance. If multiple APIs are selected, the color of lines will be used to encode the category of APIs. The consistency plot is a triangle plot where the node indicates the year and the color of the link indicate the consistency rate between two years. Consistency rate is defined as the consistency rate of prediction results in two years. The confidence chart shows the confidence changes between two years of one API. Confidence change is defined as the changes in confidence in the prediction results. The fairness chart indicates the model performance of a specific subgroup of examples. It can help users capture the performance disparity in the API.

User-defined local regime exploration. In addition to global metrics, HAPI Explorer also enables users to query particular subpopulations for detailed exploration. For example, a user may select to analyze all data points on which an ML API's prediction confidence is higher than a threshold. This is particularly useful for selecting the most accurate ML API for downstream applications relying on confidence scores. To achieve this, HAPI Explorer adopts "parallel sets", a visualization paradigm that turns a dataset into a multi-dimensional flow based on the confidence scores across time periods. Each dimension represents one year. Correlation between confidences among different years is displayed through the ribbon. Users can further click one of the ribbons in the plot, and the corresponding instances will be displayed in the following space.

Efficient individual instance search and visualization. Users can further search instances based on provided ground truth labels. The instance itself, ground truth class, confidence in different years, and the specific prediction results are displayed. Users can further check them and find intuitive insights on whether it has a significant impact on the downstream applications.

Conclusion

In this paper, we demonstrate HAPI Explorer, an interative system which enables users to understand and compare the performance of a range of ML APIs over time. To obtain fine-grained understanding, HAPI Explorer offers detailed analysis and visualization at different data hierarchy. This helps users capture the strength and weakness of ML APIs at different data regions and make appropriate decisions for different applications. We hope that HAPI Explorer can stimulate more research and exploitation of ML APIs.

References

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Chen, L.; Jin, Z.; Eyuboglu, S.; Re, C.; Zaharia, M.; and Zou, J. Y. 2022. HAPI: A Large-scale Longitudinal Dataset of Commercial ML API Predictions. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chen, L.; Koutris, P.; and Kumar, A. 2019. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the International Conference on Management of Data*, 1535–1552.

Chen, L.; Zaharia, M.; and Zou, J. 2021. How Did the Model Change? Efficiently Assessing Machine Learning API Shifts. In *International Conference on Learning Representations*.

Chen, L.; Zaharia, M.; and Zou, J. 2022. Efficient Online ML API Selection for Multi-Label Classification Tasks. In *Proceedings of the International Conference on Machine Learning*, 3716–3746.

Chen, L.; Zaharia, M.; and Zou, J. Y. 2020. Frugalml: How to use ml prediction apis more accurately and cheaply. In *Advances in Neural Information Processing Systems*, volume 33, 10685–10696.

Choo, J.; and Liu, S. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4): 84–92.

Dingen, D.; van't Veer, M.; Houthuizen, P.; Mestrom, E. H.; Korsten, E. H.; Bouwman, A. R.; and Van Wijk, J. 2018. RegressionExplorer: Interactive exploration of logistic regression models with subgroup analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 246–255.

Hosseini, H.; Xiao, B.; and Poovendran, R. 2017. Google's cloud vision api is not robust to noise. In *IEEE International Conference on Machine Learning and Applications*, 101–105.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689.

Ren, D.; Amershi, S.; Lee, B.; Suh, J.; and Williams, J. D. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics*, 23(1): 61–70.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, 601–618.