The Journal of Open Source Software

ADaPT-ML: A Data Programming Template for Machine Learning

Andrea M. Whittaker[*1]

1 University of Alberta

DOI: 10.21105/joss.04038

Software

- Review
- Repository
- Archive

Editor: Jacob Schreiber

Reviewers:

- @aaronpeikert
- @wincowgerDEV

Submitted: 02 December 2021
Published: 07 January 2022

Summary

Classification is a task that involves making a prediction about which class(es) a data point belongs to; this data point can be text, an image, audio, or can even be multimodal. This task can become intractable for many reasons, including:

- Insufficient training data to create a data-driven model; available training data may not be appropriate for the domain being studied, it may not be of the right type (e.g. only text but you want text and images), it may not have all of the categories you need, etc.
- Lack of available annotators with domain expertise, and/or resources such as time and money to label large amounts of data.
- Studying a phenomenon that changes rapidly, so what constitutes a class may change over time, making the available training data obsolete.

ADaPT-ML is a multimodal-ready MLOps system that covers the data processing, data labelling, model design, model training and optimization, and endpoint deployment, with the particular ability to adapt to classification tasks that have the aforementioned challenges. ADaPT-ML is designed to accomplish this by:

- Using `Snorkel` (Ratner et al., 2020) as the data programming framework to create large, annotated, multimodal datasets that can easily adapt to changing classification needs for training data-driven models.
- Integrating `Label Studio` (Tkachenko et al., 2020-2021) for annotating multimodal data.
- Orchestrating the Labeling Function / Label Model / End Model development, testing, and monitoring using `MLflow` (Chen et al., 2020).
- Deploying all End Models using `FastAPI` (Ramírez, 2021)

Statement of Need

Often when studying natural phenomena by creating data-driven models, processing the data becomes the largest challenge. Without a framework to build upon and implement one's ideas, researchers are forced to hastily build inflexible programs from the ground up. When

*first author

hypotheses need to be reworked or modelling a new aspect of the phenomena becomes necessary, even more time is spent on the program before finally being able to test out new ideas. This inherently causes problems, with additional problems arising such as including internal and external validation steps as an afterthought rather than a checkstop in the pipeline.

ADaPT-ML aims to be the flexible framework upon which researchers can implement their understanding of the phenomena under study. This software was created especially for any researcher with:

- Some programming experience or interest in learning how to write code based off of examples.
- Access to large amounts of unlabeled data that is constantly changing, such as social media data.
- Domain expertise or an intuition about how they would follow rules, heuristics, or use knowledge bases to annotate the unlabeled data.

ADaPT-ML takes as much of the development work as possible out of creating novel models of phenomenon for which we have well-developed theories that have yet to be applied to big data.

Related Work

An early version of this software supported the modelling of universal personal values and was complementary to the software architecture described in Gutierrez et al. (2021). During the development of this software, `Snorkel` progressed into `Snorkel Flow` (Snorkel AI, 2021), a proprietary MLOps system that incorporates data cleaning, model training and deployment, and model evaluation and monitoring into its existing data programming framework.

Acknowledgements

We acknowledge contributions from Mitacs and The Canadian Energy and Climate Nexus / Le Lien Canadien de L'Energie et du Climat for funding during the early stages of this project's development.

References

Chen, A., Chow, A., Davidson, A., DCunha, A., Ghodsi, A., Hong, S. A., Konwinski, A., Mewald, C., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Singh, A., Xie, F., Zaharia, M., Zang, R., Zheng, J., & Zumar, C. (2020). Developments in MLflow: A system to accelerate the machine learning lifecycle. *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*. https://doi.org/10.1145/3399579.3399867

Gutierrez, C. G., Whittaker, A., Patenio, K. M., Gehman, J., Lefsrud, L. M., Barbosa, D., & Stroulia, E. (2021). Analyzing and Visualizing Twitter Conversations. *CASCON x EVOKE*.

Ramírez, S. (2021). FastAPI. In *GitHub repository*. https://github.com/tiangolo/fastapi; GitHub.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *VLDB Journal*, *29*(2-3), 709–730. https://doi.org/10.1007/s00778-019-00552-1

Snorkel AI, Inc. (2021). *Snorkel Flow*. https://snorkel.ai/platform/

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., & Liubimov, N. (2020-2021). *Label Studio: Data labeling software*. https://github.com/heartexlabs/label-studio