ARISE: A Multi-Task Weak Supervision Framework for Internet Measurements

Jared Knofczynski[1,2] and Ram Durairajan[1]

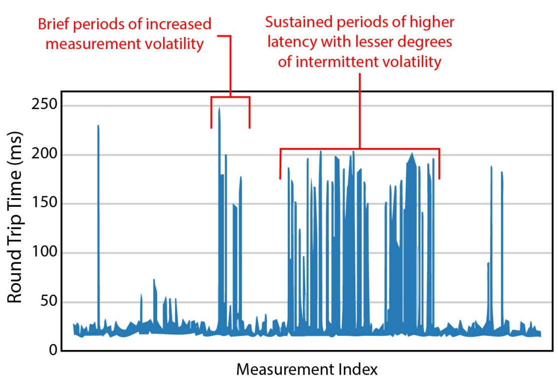
1. Computer and Information Science, UO, 2. Clark Honors College.

Introduction

- Machine learning (ML) models require large amounts of training data to perform their tasks effectively
- Network data is difficult to classify and requires domain expertise to adequately categorize
- Weak supervision allows SMEs to generate training labels programmatically using domain-specific heuristics
- Complex classification tasks require multiple ML models to perform sufficiently
- Multi-task learning (MTL) allows ML models to leverage similarities between tasks to reduce model training time and improve overall model accuracy

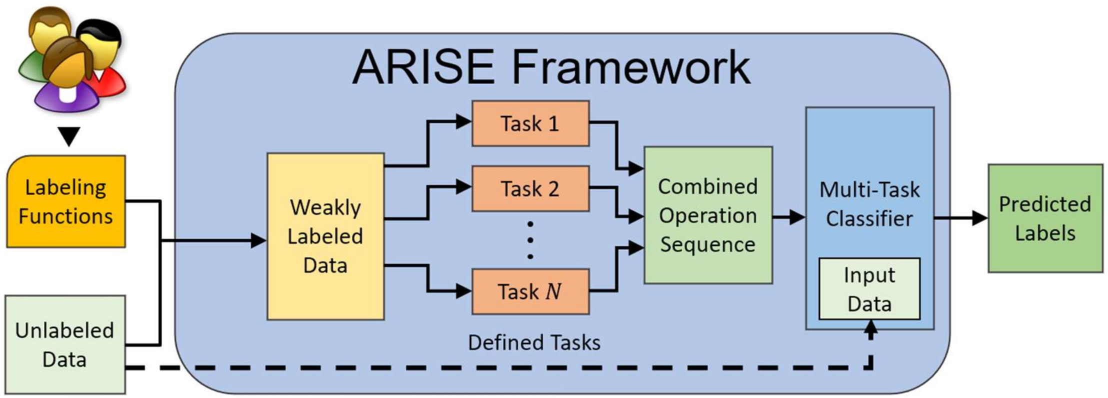
Data

We gathered 75,000 latency measurements in time series format from the Center for Applied Internet Data Analysis (CAIDA) Ark project to serve as our initial dataset, then expanded to a larger dataset containing 24.5 million latency measurements provided by the RIPE Atlas project.

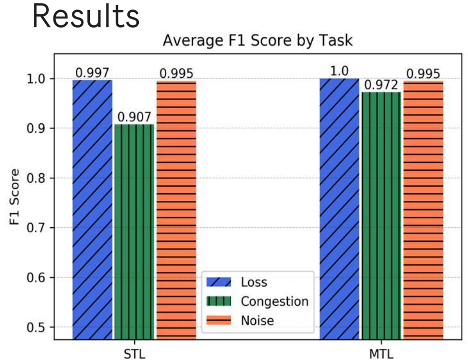


Methods

- Construct labeling functions (LFs) for noisy feature classification.
- Apply LFs to programmatically generate labels for large quantities of data.
- Develop classification tasks for distinct network features.
- Using Snorkel [1], train a multi-task classifier on these classification tasks.
- Compare model performance with previous single-task (STL) weak supervision frameworks [2].

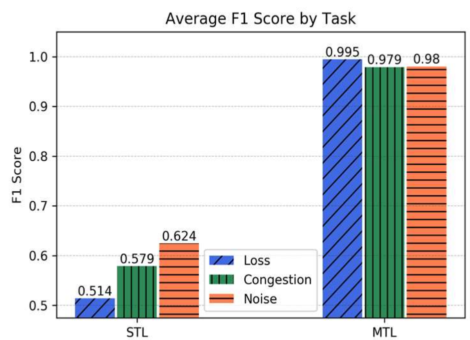


Results



Method	Task	F_1	SD	Time	SD
Single Task	LOSS	0.997	0.018	135.4s	8.22s
	CONGESTION	0.907	0.210	96.15s	11.0s
	NOISE	0.995	0.007	143.4s	8.80s
Multi Task	LOSS	1.000	0.000	15.01s	3.54s
	CONGESTION	0.972	0.016		
	NOISE	0.995	0.004		

On the CAIDA Ark datasets, multi-task learning was 2.3% more accurate and trained 96% faster than single-task learning.



Method	Task	F_1	SD	Time	SD
Single Task	LOSS	0.514	0.018	192.7s	14.0s
	NOISE	0.625	0.007	203.2s	11.0s
	CONGESTION	0.579	0.210	197.3s	8.80s
Multi Task	LOSS	0.995	0.005	34.90s	0.22s
	NOISE	0.980	0.030		
	CONGESTION	0.979	0.045		

On the larger RIPE Atlas datasets, multi-task learning was 41.2% more accurate and trained 94.1% faster than single-task learning.

Conclusions

- MTL is much faster than STL and better generalizes to larger, unseen datasets.
- Information sharing between tasks improves model classification accuracy when compared to single-task equivalents
- MTL has the potential to revolutionize how we combat network issues
- Weak supervision and MTL can be applied in other scientific fields, likely with similar levels of success.

Future Work

- Further study the performance benefits of adding additional tasks to the multi-task pipeline
- Analyze the effects of modifying the MTL information sharing capabilities to enable sharing between specific tasks, rather than all tasks.
- Examine the benefits of composing tasks in the MTL pipeline such that the predictions of one task are fed as inputs to another classification task.

References

[1] A. Ratner, B. Hancock, J. Dunnmon, R. Goldam, and C. Ré, "Snorkel MeTaL: Weak Supervision for Multi-Task Learning," *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 2018.

[2] Y. Lavinia, R. Durairajan, R. Rejaie, and W. Willinger, "Challenges in Using ML for Networking Research: How to Label If You Must," 2020.

UNIVERSITY OF OREGON CAS NSF ripple

Many thanks to Professor Durairajan, Ripple, and the National Science Foundation for their continued support.

May 2021