JULY 12–16
2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

Data Shapely based Auto-labeling algorithm

Ti Bai, Brandon Wang, Biling Wang, Dan Nguyen and Steve Jiang
Medical Artificial Intelligence and Automation (MAIA) Laboratory,
Department of Radiation Oncology,
UT Southwestern Medical Center, Dallas, Texas
Ti.Bai@UTSouthwestern.edu

MAIA Laboratory
UT Southwestern Medical Center
Radiation Oncology

INTRODUCTION

- Training deep neural networks for health care tasks requires a large amount of correctly labeled data
- Manually annotation requires highly expertise domain knowledge, and hence is very expensive and time-consuming in medical domain
- Little research efforts have been devoted

AIM

- In this study, we developed an automated method to label data using the data Shapley algorithm.

METHODS

- Given a dataset D, the data value of data point i can be characterized by its Data Shapley [2] as following:

$$\phi = C \sum_{S \subset D - \{i\}} \frac{L(S \cup \{i\}) - L(S)}{\binom{n-1}{|S|}}$$

where L is the associated loss function

- Suppose we have a small amount of well-labeled dataset, and a large amount of unlabeled dataset.

1) Multiple copies of unlabeled dataset are generated with different pseudo labels
2) Combine with above labeled dataset

METHODS (CONTINUED)

3) One epoch of training from scratch, and record the relative loss change
4) Repeated 3) by 1000 times, the averaged loss change can be considered as the negative Shapley value.
5) The copy with largest Shapley value indicates the correct label
6) Repeat this procedure for many times, a large amount of unlabeled dataset can be freely labeled by using only a small amount of well-labeled dataset.

DATASETS

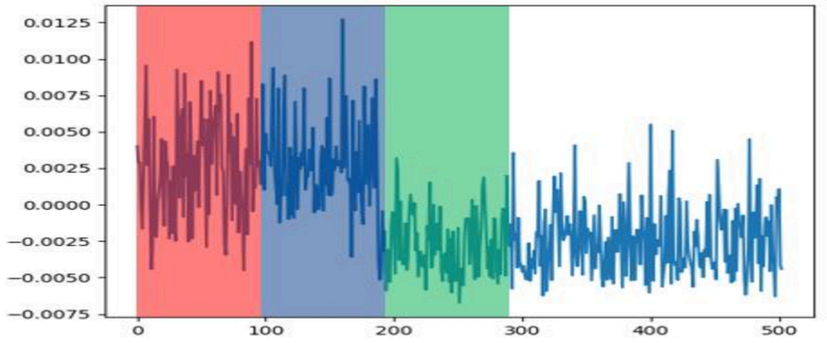
- The Osteosarcoma Tumor (OSTU) Identification Dataset with 344 images/3 labels was used.
- The OSTU datasets are split into 220(labeled)/30(test)/94(unlabeled) datasets.
- To explore the performance dependence on the size of the labeled dataset, different amount of labeled images are separately used as the labeled dataset.

CONCLUSIONS

- An auto-labeling algorithm was developed with more than 90% top-1 accuracy by using the data Shapley value.
- The auto-labeling performance can be further improved as we added more labeled images.

RESULTS

SHAPLEY VALUE OF DIFFERENT DATA POINTS



- The shading region represents unlabeled datasets.
- Green are correct label; the other two colors have wrong label copies.

EFFECT OF LABELED DATASET SIZE ON THE AUTO-LABELING ACCURACY

Labeled dataset size	0	31	63	94	126	157	189	220
Top-1 accuracy	69%	74%	72%	91%	96%	90%	98%	95%

REFERENCES

- *Shapley, L. S., Roth, A. E. et al. The Shapley value: essays in honor ofLloyd S. Shapley (Cambridge University Press, 1988).*
- *Ghorbani, A. and J. Y. Zou (2019). "What is your data worth? Equitable Valuation of Data." arXiv 1904.02868v2.*