17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Estimating Contribution of Training Datasets using Shapley Values in Data-scale for Visual Recognition

Takayuki Semitsu[1*] Mitsuki Nakamura[1] Shotaro Ishigami[1] Toru Aoki[1]

Teng-yok Lee[1] Yoshimi Isu[1]

Abstract

In this paper, we propose a method to measure contributions of multiple datasets i.e. how much a specific dataset contributes to improve accuracy of the model. Our method is based on shapley value, of which purpose is to measure contribution by difference of the accuracy of the models. Unlike previous method, our method first converts the accuracy to data-scale measurements using fitted log curve. We calculate contributions in a fair way that each trials are evaluated not by its improvements of accuracy, but by the number of data needed to make the improvements. Our method can avoid overestimation of contributions in small data cases. To evaluate the proposed method, we trained models for Person Re-Identification tasks with combinations of datasets, and calculated contributions of each datasets. Results show that the proposed metrics can effectively reduce the overestimations in small data cases, while the contributions maintain good properties such as local accuracy and additive law derived from shapley value definition. We also proposed normalization of shapley values in data-scale by its actual number of instances, which indicates intrinsic importance of a dataset per instance.

1 Introduction

Various visual recognition applications (e.g. Object Detection and Person Re-Identification) are studied in the field of both research and industry. Collecting training data is critical for developing models for each task. Even for same the application, users often need to collect different datasets for different target scenes in order to learn scene-specific model. Combining multiple datasets, user can get a model with better robustness and accuracy, compared to the one trained on a single dataset. In this case, measuring contributions of each dataset, i.e. how much individual datasets contribute to the improvements in training, is important. With the measured contributions, we can propose better data collection policy for additional data, or acquire efficient data subsets achieving higher accuracy with less data given specified number of samples.

Given all training datasets combined, one way to measure contribution of a training dataset w.r.t. the target scene is to train two models with and without the specific dataset and take the difference. We call it as "last-one-mile gain." Intuition behind this metrics is that the improvement of 1% is more important when the dataset is larger. The problem of this baseline metrics is that the same values of difference may means different contributions. As a example, we assume that the specific training dataset is already included in the base datasets. In this case, the contribution (i.e. difference of accuracy) is likely to be small since the actual training samples are identical. In this way, the metrics highly depends on the presence of specific training samples, even though it could vary in practice.

In this paper we propose a method to quantitatively measure the contributions of dataset by using Shapley values. We pointed out that calculating Shapley values simply with model accuracies could overestimate the contribution in the case of small data size. We applied accuracy-to-size conversion based on log-curve.

Our contributions are:

1. We established a novel method to measure contribution of each datasets with source dataset fixed, in which contributions are calculated in a fair way that each trials are evaluated not by its improvements of accuracy, but by the number of data needed to make the improvements.

2. We showed the effectiveness of our approach through the experiments in the field of Person Re-Identification task, with 5 public datasets. We confirmed that our method can effectively reduce the overestimations in small data cases, while the contributions have good properties such as local accuracy and additive law derived from Shapley value definition.

2 Related Works

Shapley Value and its approximation: Shapley Value (SV) [9], which is first introduced in Game Theory, aims to fairly measure contributions of each players in a quest. Since original SV requires $O(2^N)$ trials for N datasets, several approximation methods such as Monte-Carlo sampling-based approach[1] are proposed.

Shapley Value in Machine Learning: In machine learning literature, local explanation to out-

*[1]Mitsubishi Electric Corporation, 5-1-1, Ofuna, Kamakura, Japan

puts of machine learning model can be calculated as SV of a conditional expectation function of the model [8].

Model Accuracy and Data Size: Scaling laws between model accuracy and data is reported for vision tasks [10] and natural language tasks [5]. Accuracy (e.g. mAP) increases logarithmically as data increases, or test loss decreases by $e^{\frac{1}{n}}$ as data n increases.

Dataset Shapley: Several works focuses on measuring contributions of data instances (or group of instances) included in a dataset using SV[4, 2]. These works focused on efficient approximation of SV calculation. On the other hand, we try to revisit what SV really compares, and pointed out the problem of comparing contributions in linear-scale accuracy measurement. We propose the method to calculate SV in log-scale data measurement (we refer it as "data-scale"), which is not mentioned in previous works.

3 Method

In this paper, we propose a method to calculate quantitative contributions of individual dataset to its model accuracy, given combinations of all the datasets and specific target scene.

Suppose we have multiple datasets $D = \{D^{\text{train}}, D^{\text{test}}\}$ with train-sets denoted by i (i.e. $D^{\text{train}} = \{D_i^{\text{train}} | i = 1, ..., N\}$) and test-sets denoted by j (i.e. $D^{\text{test}} = \{D_j^{\text{test}} | j = 1, ..., N\}$). P corresponds to all possible combinations of train-sets D^{train}. Thus $|P| = 2^N$. The standard Shapley value ϕ_i of train-set D_i^{train} with test-set D_j^{test} (i.e. target scene) is calculated as:

$$\phi_i(P, v_j) := \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} \left(v_j(S \cup \{i\}) - v_j(S)\right) \tag{1}$$

where $v_j(X)$ is a function which returns the reward of the model trained on X and evaluated on test-set j. Note that function $v_j(x)$ depends on test set j, since contribution should depend on test-set.

One simple choice for v is to take the accuracy of the model. However, this approach has two problems. First, there are huge accuracy gap between models with and without source-set, which is train-set of test-set j (=target-scene), as shown in figure 1.

Second, a model with smaller data tends to have larger improvements given same amount of additional data (See figure 2 and 3). Taking the difference of accuracy as measurement for contribution, Shapley value would not differentiate these differences. In this case, 1% improvements at small data gets same contribution as the one at large data, which is fundamentally different.

As for the first problem, we fix the presence of source-set in each experiment. We defined two

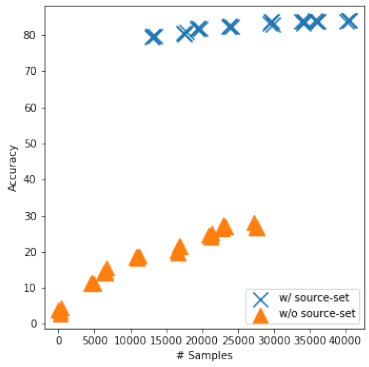


Figure 1: Plots of accuracy and # samples. Each point represents a evaluation result of a model trained on all possible combinations of 6 train-sets described in section 4, and evaluated on Market-1501. Presence of source-set (i.e. Market-1501 train-set) makes huge performance gap.

groups: the first group includes models trained with the source-set. The other group includes models trained without the source-set. With P' which corresponds to all possible combinations of $D^{\text{train}} \setminus \{j\}$ and $S \subseteq P'$, the training dataset of the first group is denoted as $T^+ = S \cup \{j\}$, and the training dataset of the second group is denoted as $T^- = S$

Second, we propose to measure contributions in data-scale. As reported in [10], the relationship between accuracy and data size can be described by log-curve. We convert the accuracy to data scale measurement by following procedure. We have $M(= |P'|)$ results $\{s_k^j, a_k^j\}(k = 1, ...M)$, where s_k^j is total number of images in the train-sets and a_k^j is accuracy of the trained model evaluated on test-set j.Based on the empirical results about scaling laws of model [2, 10], we describe the relationship between accuracy a and the number of training samples s of individual model as:

$$a = f_j(s) = \alpha_j(\log s - \log \beta_j), \tag{2}$$

where α_j and β_j are coefficients for the log curve to optimize. Note that this corresponds to linear line of log scale s. Given training results $\{a_k^j, s_k^j\}$, we fit equation 2 to the data using least-square method.

Using fitted f_j, Shapley value of train-set i, given test-set j, and with data-scale measurement can be calculated as:

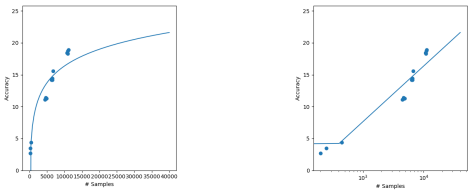
$$\psi_i(P', v_j) := \sum_{S \subseteq P' \setminus \{i\}} \frac{|S|!(|P'| - |S| - 1)!}{|P'|!} \left(f_j^{-1}(v_j(S \cup \{i\})) - f_j^{-1}(v_j(S))\right) \tag{3}$$

4 Experimental Results

We evaluated the effectiveness of the proposed method using datasets of Person Re-Identification(Re-ID) task, which aims to find

Table 1: Number of training images for each datasets. $(\cdot)$ denotes abbreviated characters used in other experiments.

Dataset Names	# Images	# IDs
Market1501 (m)	12936	751
GRID (g)	250	125
PRID (p)	200	100
SenseReID (s)	4428	1718
CUHK02 (c)	6308	1577



(a) linear-scale plots (b) log-scale plots

Figure 2: Evaluation results of OSNet on market1501. Accuracies and size for all possible combinations of datasets are plotted with linear-scale (left) and data-scale(log-scale) (right).

query person from other gallery cameras. We trained models on all combinations of datasets and calculated the contributions of each datasets. Note that our methodology is applicable to any other tasks as long as scaling laws for model holds. [10] and [5] reported scaling law for classification task. Re-ID task can be regarded as classification task given datasets to evaluate. In this section, we showed that: (i) the data is effectively modeled by the scaling law (i.e. log-curve), on which Shapley values of each dataset can be calculated (See section 4.1). and (ii) the calculated Shapley values can be normalized in several way, providing different insights about training results.

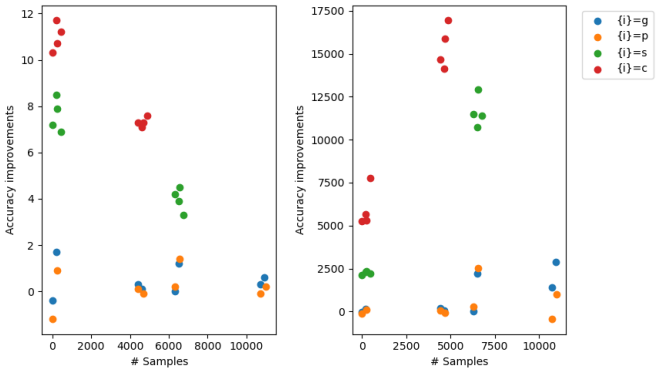
4.1 Shapley Values with Data-scale Measurement.

We used 5 datasets for evaluation: Market1501 [12], GRID [7], prid2011 [3], CUHK02 [6], and SenseReID [11]. We used OSNet [14] as models to measure accuracies. Table 1 shows overview of each datasets. Our implementation is based on torchreid [13].

At each experiment, we fix the source-set (i.e. train-set which is from same dataset as test-set) included or not included in the train-sets. Next, we trained 2^N $(N = 5)$ models for all possible subsets of datasets. Contributions are calculated using these results.

We defined two baselines. One is last-one-mile gain described in section 3. It indicates the contribution of the specific train-set, given all other train-sets included. The other is Shapley value of datasets, which calculates contributions by the difference of accuracy, whereas our method converts the accuracy to data-scale measurement.

Table 2 summarizes the comparison of the properties of methods. Shapley-based approaches have good properties such as local accuracy and additive law, derived from definition of SV. Local accuracy is a property that summation of contributions corresponds to the function value (i.e. accuracy of the model). Additive law is the property that if c_1 and c_2 are contributions of S_1 and S_2, contribution of $S_1 + S_2$ is $c_1 + c_2$. Only the proposed method takes data size into considerations of contribution.



(a) linear-scale plots (b) log-scale plots

Figure 3: Plots of $(|S|, \Delta v_{ji})$, where $\Delta v_{ji} = v_j(S) - v_j(S/\{i\})$ in (a), and $\Delta v_{ji} = f^{-1}(v_j(S)) - f^{-1}(v_j(S/\{i\}))$ in (b). Models are evaluated on Market-1501 and source-set is not included in the train-sets. From (a), we can see that there are strong negative correlation between the number of images and the amount improvements given train-sets. As shown in (b), accuracy-to-data-scale conversion reduces this correlation. Best viewed in color.

Table 2: Comparison of contribution metrics.

Properties	Last-one-mile gain	SV in linear-scale	SV in data-scale
local accuracy?	No	Yes	Yes
additive law?	No	Yes	Yes
data-size aware?	No	No	Yes
# trainings	$N+1$	2^N	2^N

Note that in this paper we calculate exact Shapley value to evaluate the proposed methodology. Approximation such as Monte-Carlo sampling-based method [1] can be applied in practice.

Figure 4 shows the visualization of Shapley values calculated on Market-1501. As shown in figure 3(a), simply taking difference of accuracy leads to overestimation of results with small training samples. Figure 2(b) shows that the relationship between the accuracies and the data size can be described by log-curve.

Table 3 shows the confusion matrix of Shapley values. Each column is a result for specific test-set denoted in the first row.

Table 4 shows the calculated contributions by 3 different methods last-one-mile gain normalized to data-scale, SV in linear scale, and SV in data-

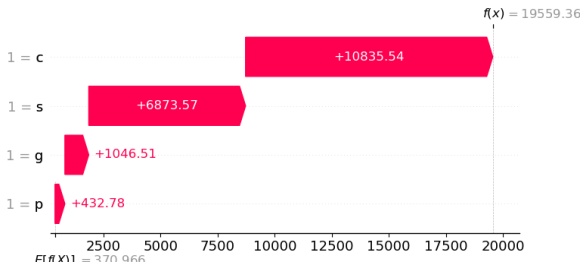


Figure 4: Visualizations of Shapley values. Red arrow shows positive contributions and blue arrow shows negative contributions. Sum of all Shapley values corresponds to the approximate dataset size of the model trained on all datasets.

Table 3: Confusion matrix of SV in data-scale. Column of "train-set" represents train-set to calculate contributions. Column of "test-set" shows data-scale Shapley values for each test-set. Last column and row show summation of the row and column respectively. Note that we only use SenseReID for training.

train-set	test-set				sum
	m	g	p	c	
m	0	26000	388400	19900	434300
g	1000	0	272000	4600	277700
p	400	15700	0	2600	18700
s	6900	-19800	283700	6400	277200
c	10800	1500	359500	0	371900
base	400	800	100	300	-
sum	19600	24200	1303800	33700	-

scale.

4.2 Comparison using Normalized Shapley Values

As described in section 4.1, accuracy can be converted to its dataset size with logarithmical relationship. By converting accuracy-scale values to data-scale values, we can fairly compare dataset contributions as amount of essential data size added. By the definition of SV, sum of dataset contributions given specific test-set corresponds to the actual data size. We introduce several ways to normalize the contributions, and how to interpret the values.

Average Importance per Instance: Table 5 shows average importance per instance, which is contributions of datasets given same amount of data. As shown in table 5, some datasets have large positive contributions per instance (e.g. (c)uhk02)), whereas some dataset have zero or large negative contribution per instance (e.g. (g)rid and (p)rid).

5 Conclusion

In this paper we proposed we propose a method to measure contributions of datasets using Shapley values. Based on the empirical results about scaling law of models, we learn a function to convert accuracy-scale values to data-scale ones. With this conversion, contributions are calculated in a fair way that each trial are evaluated not by its improvements of accuracy, but by the number of data needed to make the improvements. We showed the effectiveness of our method by the experiments on Person Re-Identification task, using 5 public datasets.

Table 4: Contributions using three different approaches (i.e. Last-one-mile gain normalized to data-scale, SV in linear scale, and SV in data-scale). Target-set is Market-1501 and all models are trained without source-set. LOM exaggerated the single case results, whereas SV in linear scales are inappropriately smoothed since every datasets provides larger contributions when S is small. The proposed method provides reasonable results in between the former two, with good properties mentioned in table 2.

source	Last-one-mile gain	SV in linear-scale	SV in data-scale
g	0.60	0.35	1.01
p	0.20	-0.05	0.42
s	3.30	5.62	6.64
c	7.60	9.08	10.47
base	0.00	3.90	0.36

Table 5: Per-instance contributions of each datasets. The proposed SV in data-scale is normalized by their actual data size. Some datasets have large positive contributions per-instance (e.g. (c)uhk02)), whereas some dataset have large negative contribution per-instance (e.g. (p)rid).

	SV in data-scale	# Images	SV / # Images
g	1047	250	4.19
p	433	200	2.16
s	6874	4428	1.55
c	10836	6308	1.72

References

[1] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 12 2013.

[2] A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251, 2019.

[3] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[4] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes,

N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[5] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[6] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[7] C. Liu, S. Gong, C. C. Loy, and X. Lin. Evaluating feature importance for re-identification. In *Person re-identification*, pages 203–228. Springer, 2014.

[8] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[9] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[10] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[11] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[13] K. Zhou and T. Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019.

[14] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.