KDD2024
BARCELONA, SPAIN

Approximating Memorization Using Loss Surface Geometry for Dataset Pruning and Summarization

Andrea Agiollo[1], *Kim Young In*[2], and Rajiv Khanna[2]

[1]University of Bologna
[2]Purdue University

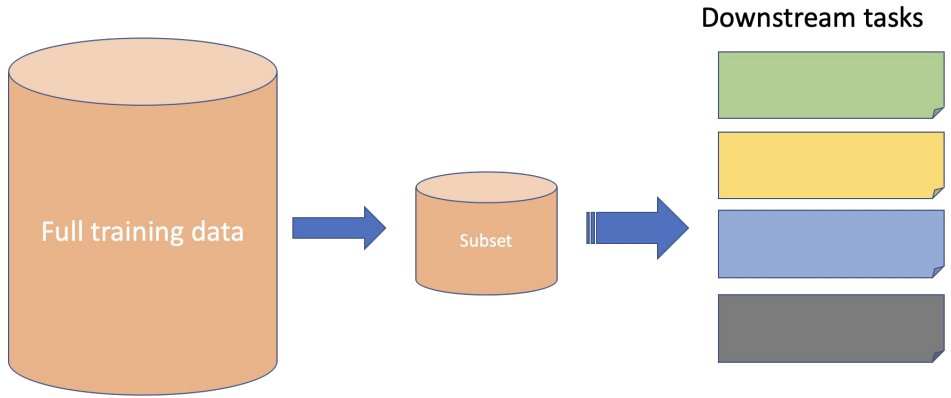
PURDUE UNIVERSITY | Department of Computer Science

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Data summarization

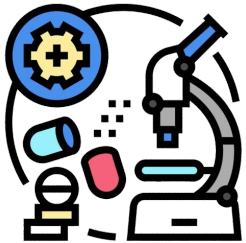
Goal: Compress data used for training with minimal loss in predictive performance

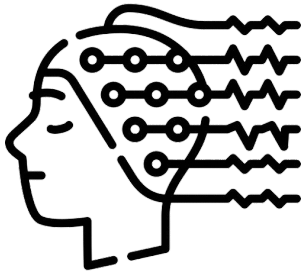
- Key**:** Find informative example data points
- Retain data interpretability: avoid data transformation
- One-shot summarization



Data summarization applications

- *Drug discovery*: searching through over 10^{60} molecules and compressing the data by selecting fewer molecules that are predictive of a certain property accelerate the drug discovery task
- *Neuroscience*: electroencephalogram signal analysis require processing of billions data points
- *Metagenomics*: learning characteristics of genetic clinical material where number of available combinations for human DNA has approximately order of 10^{120}
- *High energy physics:* CERN's Large Hadron Collider generates close to 1 million petabytes of data for each year



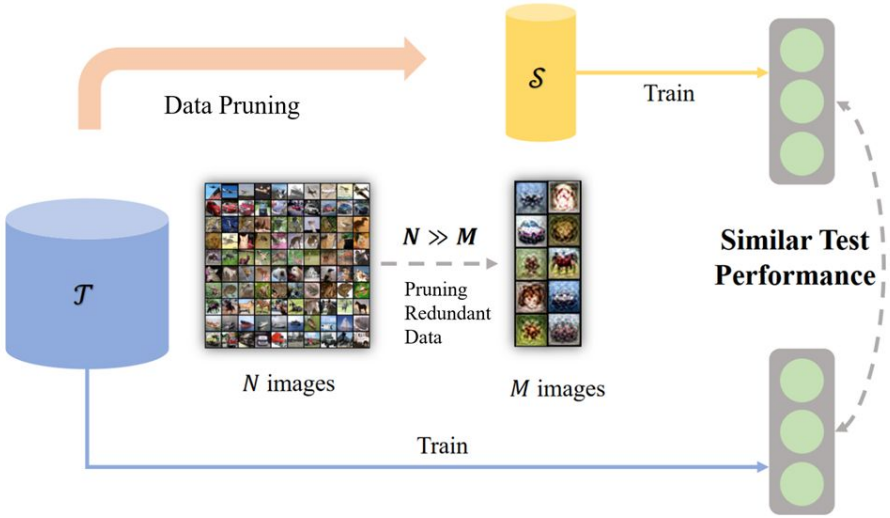






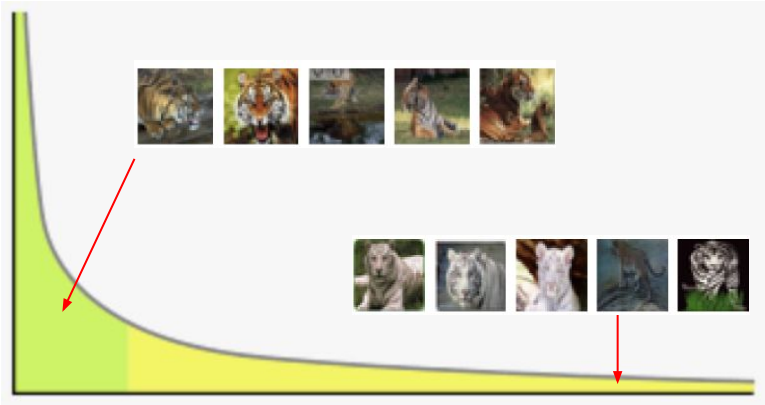
Data summarization state-of-the-art

- **Key:** Find informative example data points for NN training
- Proxies for importance of a data point:
  - Ease of learning: How soon is a data point learnt?
  - Forgetfulness: Once it is learnt, is it ever “forgotten”?
  - Realtime pruning: Keep a validation set and identify during training which data points can be “pruned away”
  - Static pruning: Use interpretability techniques (e.g. Influence functions or Fisher Kernels) to identify which training data points are important for the validation set



Real world data distributions

- Composed of **typical** subpatterns and numerous **atypical** subpatterns
- For close-to-optimal performance, model must identify atypical subpatterns as well
- Identification of the subpatterns lies in a spectrum between perfect **learning** and perfect **memorization**



Memorization scores

- **Our approach is based on memorization scores**
- Memorization score measures self-influence a data point has on itself, quantifying where the data point lies in the learning-memorization spectrum (Feldman and Zhang, 2020)

$$m(\mathbf{x}_i) = P\left[h_{\mathcal{T}}^{\mathcal{A}}(\mathbf{x}_i) = y_i\right] - P\left[h_{\mathcal{T}^{\setminus i}}^{\mathcal{A}}(\mathbf{x}_i) = y_i\right]$$

- Low memorization score: Removing from training set has minimal effect as model **learned** subpatterns from numerous other similar data points → Typical

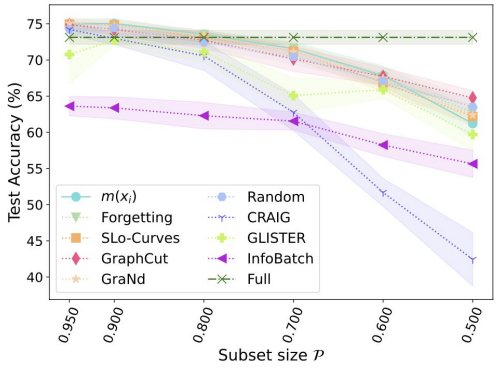
- High memorization score: Removing from training set has significant effect as the model **memorized** the subpatterns solely from itself (Singleton) or along with few other samples → Atypical

Memorization scores for data summarization

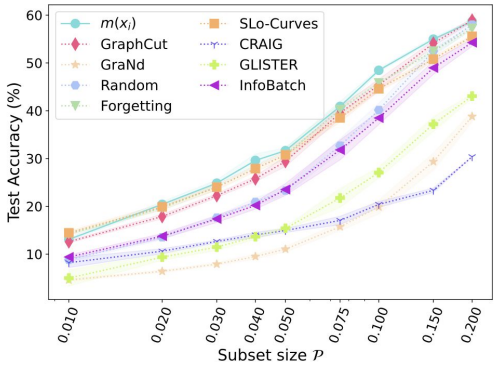
- Want to retain typical samples that generalize well → Priority for small subset
- Don’t want too many typical samples as they contain redundant information → Large subset
- Want to include fair amount of atypical samples to not miss out on rare subclasses → Large subset

Memorization scores for data summarization

- Sample high mem-score data for data pruning (large subset)
- Sample low mem-score data for few-shot data summarization (small subset)



(a) Data pruning

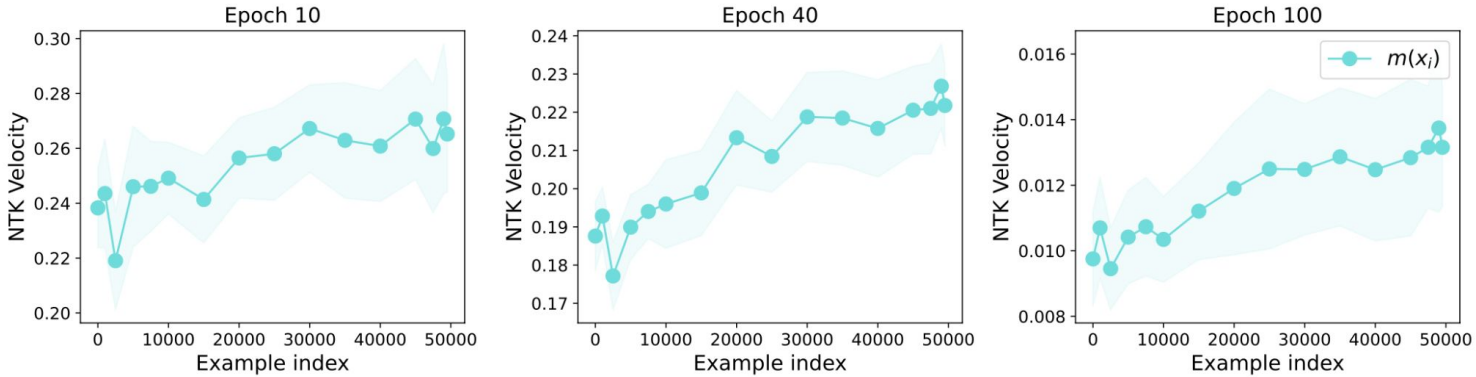


(b) Few-shot data summarization

Finding #1: **Memorization scores are very effective for data summarization.**

Why are memorization scores effective?

- Analyse evolution of Neural Tangent Kernel (NTK) over memorization scores
- Low memorization samples learnt quickly and early
- High memorization samples harder to learn and learnt later (if ever)



Finding #2: **NTK velocity directly proportional to memorization.**

Issues with memorization scores

- Exact memorization score computation is computationally infeasible
- Even with sampling strategy, requires training 4000 NN models to compute memorization scores over CIFAR100
- Use influence functions? → found to be fragile for deep NN

Challenge: tractably approximate memorization scores

Intuition #1: tackle data typicality/importance problem rather than exact memorization

Intuition #2: leverage geometric property of loss landscape to effectively and efficiently identify sample typicality/importance and use it for data summarization

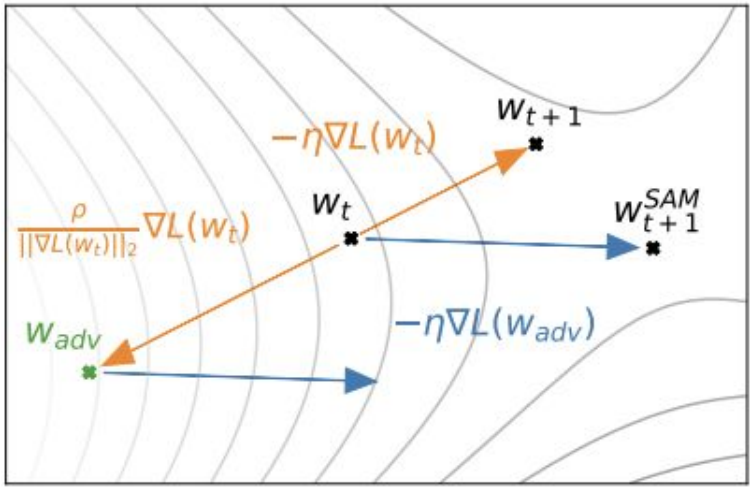
Sharpness Aware Minimization (SAM)

- Optimization algorithm designed to find a flatter minima

$$\min_{\boldsymbol{w}} L_S^{SAM}(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_2^2 \quad \text{where} \quad L_S^{SAM}(\boldsymbol{w}) \triangleq \max_{||\boldsymbol{\epsilon}||_p \leq \rho} L_S(\boldsymbol{w} + \boldsymbol{\epsilon})$$

- Two steps:
  - Gradient ascent to maximum loss in the vicinity of the current minima
  - Gradient descent

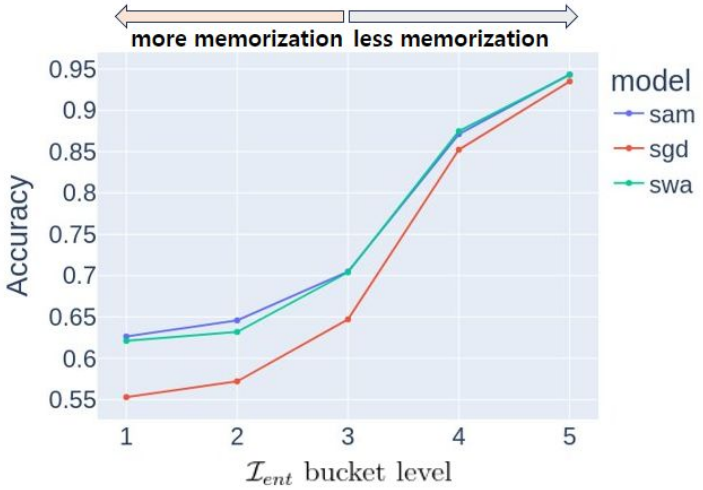
- Empirically shown performance gains compared to standard SGD



(Foret et al., 2020)

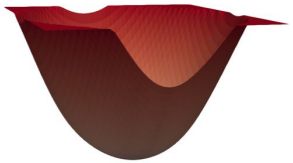
SAM vs SGD

- Previous work dissects performance gains of SAM and finds that the gains come from atypical samples
- Employed a new metric for measuring how much a test sample depends on memorization → data typicality
- We can **reverse-engineer** this **discrepancy** to approximate data typicality and memorization

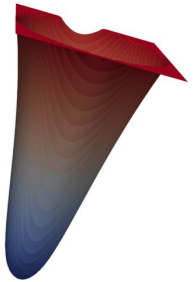


(Kim et al., 2023)

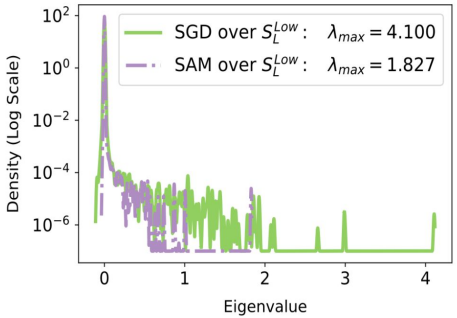
Further exploration



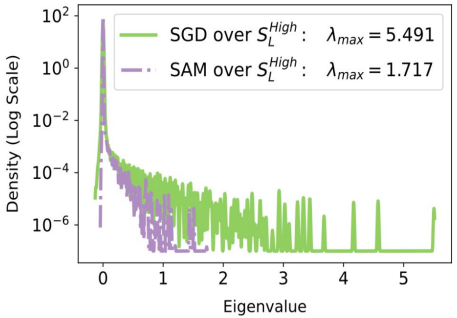
(a) Low mem



(b) High mem



(a) Low memorization



(b) High memorization

Finding #3: **typical (low memorization) data points trigger flatter minima by default.**

Finding #4: **gap between SAM and SGD is greater for high memorization samples.**

SAMIS

Intuition: leverage discrepancy between SAM and SGD over data points to measure data point typicality/memorization

- SAMIS-L: absolute difference in the *loss* between the SAM model and SGD model (precise in supervised settings):

$$S_L(\mathbf{x}_i) = |\mathcal{L}\left(h^{SAM}_{S^{\setminus i,..,z}}(\mathbf{x}_i), y_i\right) - \mathcal{L}\left(h^{SGD}_{S^{\setminus i,..,z}}(\mathbf{x}_i), y_i\right)|$$

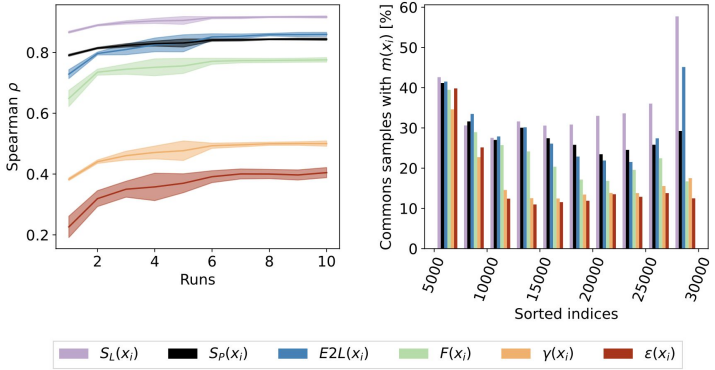
- SAMIS-P: absolute difference in the *prediction probability* between the SAM model and SGD model (useful in unsupervised settings):

$$S_P(\mathbf{x}_i) = \frac{\sum_{c=1}^{C} |P\left(h^{SAM}_{S^{\setminus i,..,z}}(\mathbf{x}_i) = c\right) - P\left(h^{SGD}_{S^{\setminus i,..,z}}(\mathbf{x}_i) = c\right)|}{C}$$

SAMIS & memorization

How well does SAMIS approximate memorization scores on CIFAR100:

- SAMIS-L reaches highest spearman correlation up to 0.92
- SAMIS correlates well both on low and high memorization extremes
- After visualization, several images are matching between memorization scores and SAMIS
- Similar results with other non flatness-aware optimizers (Adam, RMSProp, etc.)





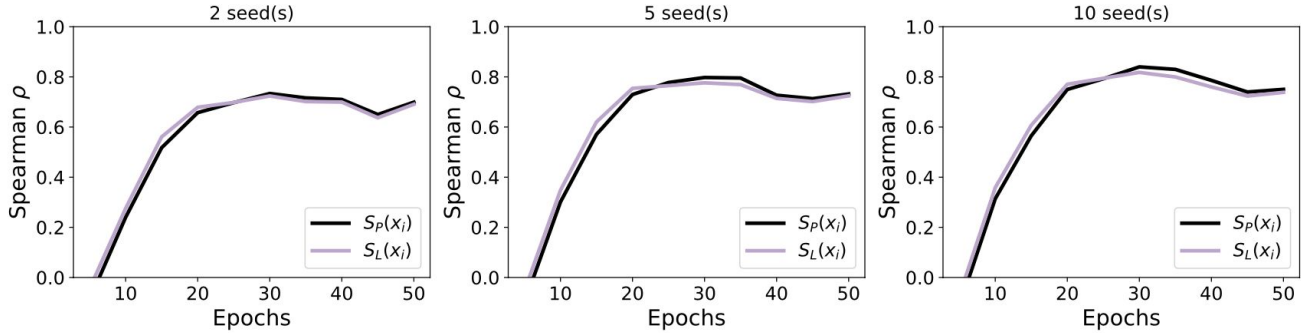
(a) Low memorization/proxy



(b) High memorization/proxy

SAMIS efficiency

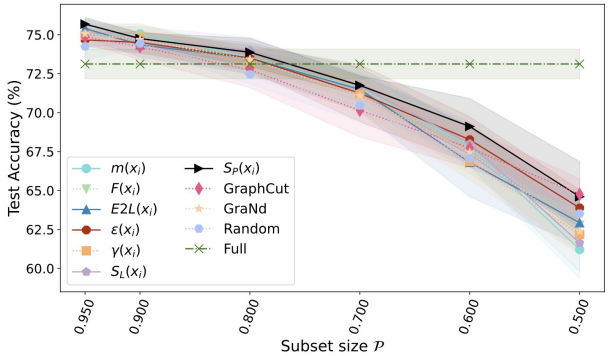
- SAMIS requires training $2nk$ models: n number of seeds and k number of dataset splits
- n=5 and k=10 highest correlation and 40× speed up compared to 4000 NNs for memorization
- Alternative version of SAMIS using premature training: 0.8 correlation and 2000× speed-up

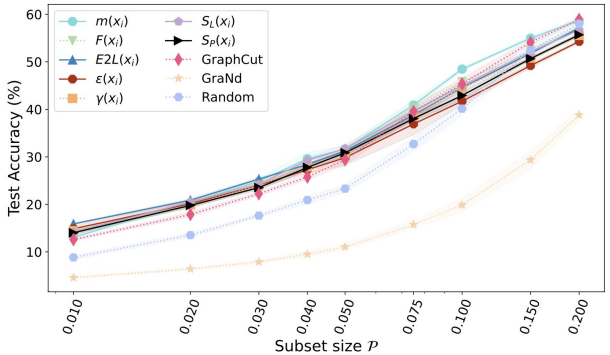


Finding #5: **SAMIS can effectively and efficiently approximate memorization.**

SAMIS for dataset summarization

- Experiments on CIFAR10, CIFAR100, and SVHN datasets
- Compare SAMIS-based subset selection against 7 well-known data summarization baselines
- Consider both a data pruning and a few-shot learning setup





Finding #6: **SAMIS-based sampling outperforms state-of-the-art data summarization over most sampling ratios.**

More SAMIS applications

- SAMIS tested on vision tasks only, but general enough to be applicable to any data structure
- Interesting directions:
  - *Open-set active learning*: SAMIS scores to be used to identify the most relevant samples to be labelled → to be released on arxiv!
  - *Curriculum learning*: ML models trained in a meaningful order, from the easy/typical samples to the hard/atypical ones
  - *Federated learning*: SAMIS can balance the complexity of the learning task of each client, depending on its resource availability, using data typicality
  - *Knowledge distillation*: SAMIS used to balance how much the student model should adhere to the teacher's predictions depending on the data typicality
  - More suggestions are welcome!

Conclusions

- Memorization scores can be used for data summarization effectively
- NTK velocity is proportional to memorization scores of data samples and it can be used to identify strong data summarization subsets
- Clean typical samples (low memorization) trigger flatter loss minima compared to atypical, rare or noisy samples (high memorization)
- Leveraging the discrepancy between SAM and SGD can effectively approximate memorization scores
- SAMIS approximates very closely memorization while achieving up to 2000x speed-up in computation
- High fidelity memorization proxies attain similar performance of memorization scores for data summarization, surpassing state-of-the-art solutions while also being scalable

THANK YOU

Full paper:

Source code:

References

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-aware minimization for efficiently improving generalization,” *International Conference on Learning Representations*, May 2021, [Online]. Available: https://openreview.net/pdf?id=6Tm1mposlrM

V. Feldman and C. Zhang, “What neural networks memorize and why: Discovering the long tail via influence estimation,” *Neural Information Processing Systems*, vol. 33, pp. 2881–2891, Aug. 2020, [Online]. Available: https://papers.neurips.cc/paper/2020/file/1e14bfe2714193e7af5abc64ecbd6b46-Paper.pdf

Y. I. Kim, P. Agrawal, J. Royset, and R. Khanna, “On Memorization and Privacy Risks of Sharpness Aware Minimization,” *Arxiv*, 2023. https://arxiv.org/abs/2310.00488

Brief photo caption. Place in top left or right corner. Acumin Pro Bold 18 pt. Make text black or white for legibility.

123

TOPIC OR TITLE

Fact or highlight. Acumin Pro Medium 24 pt.
Keep it short with bite-size chunks of information.

KDD2024
BARCELONA, SPAIN

Approximating Memorization Using Loss Surface Geometry for Dataset Pruning and Summarization

Andrea Agiollo[1], Kim Young In[2], and Rajiv Khanna[2]

[1]University of Bologna
[2]Purdue University

PURDUE UNIVERSITY | Department of Computer Science

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

KDD2024
BARCELONA, SPAIN

Approximating Memorization Using Loss Surface Geometry for Dataset Pruning and Summarization

Andrea Agiollo[1], Kim Young In[2], and Rajiv Khanna[2]

[1]University of Bologna
[2]Purdue University

PURDUE UNIVERSITY | Department of Computer Science

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Approximating Memorization Using Loss Surface Geometry for Dataset Pruning and Summarization

Andrea Agiollo[1], Kim Young In[2], and Rajiv Khanna[2]

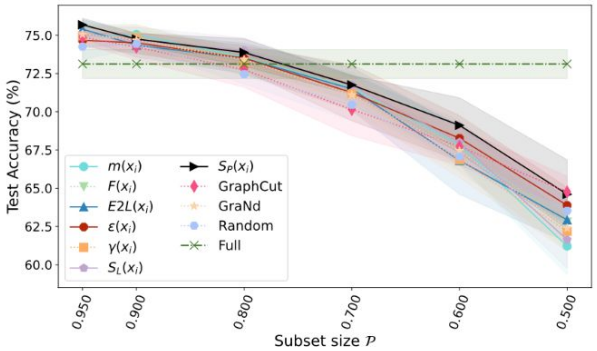
[1]University of Bologna
[2]Purdue University

PURDUE UNIVERSITY | Department of Computer Science

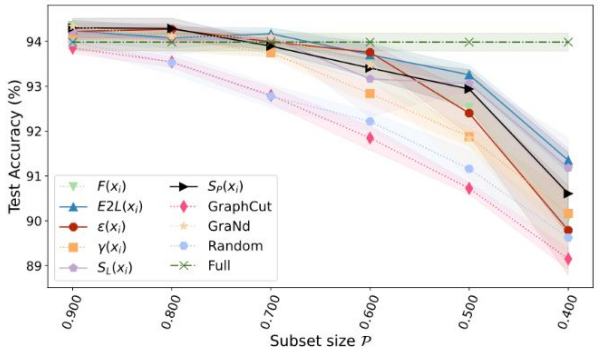
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

SAMIS for dataset summarization (Part I)

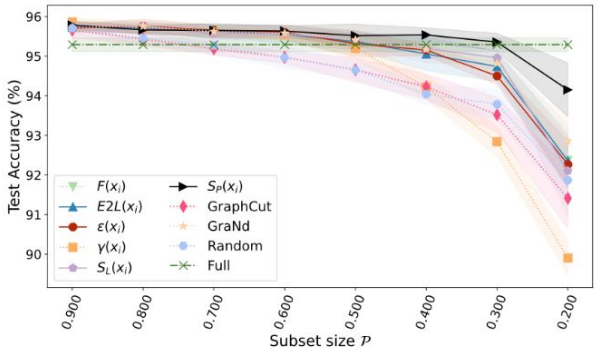
- Experiments on CIFAR10, CIFAR100, and SVHN datasets
- Compare SAMIS-based subset selection against 7 well-known data summarization baselines
- Consider both a data pruning and a few-shot learning setup
- Following finding #2, sample highest SAMIS for data pruning (large subset) and lowest SAMIS for few-shot summarization (small subset)



(a) CIFAR100



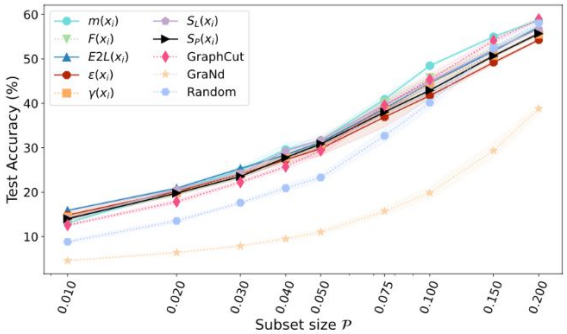
(b) CIFAR10



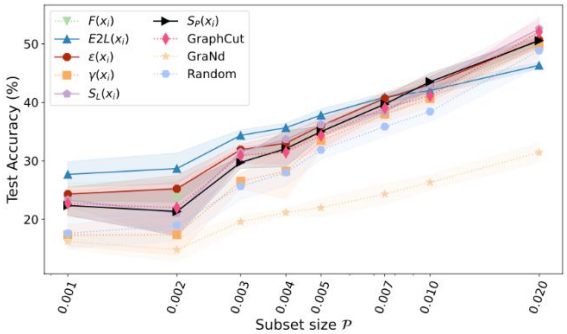
(c) SVHN

SAMIS for dataset summarization (Part II)

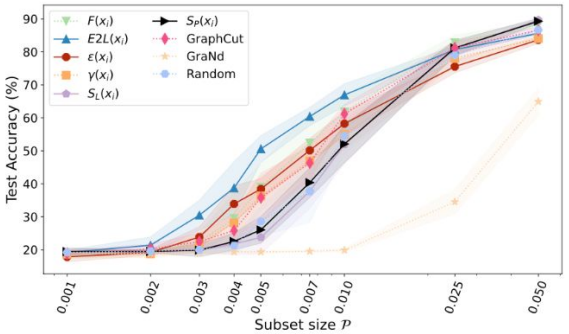
- Tie between clean samples having lowest memorization and SAMIS make memorization proxies perform less well for smaller sampling ratios



(a) CIFAR100



(b) CIFAR10



(c) SVHN

Finding #6: **SAMIS-based sampling outperforms state-of-the-art data summarization over most sampling ratios.**