PURDUE UNIVERSITY

U.S. National Science Foundation

Collins Aerospace

CoCoSys
CENTER FOR THE CO-DESIGN OF COGNITIVE SYSTEMS

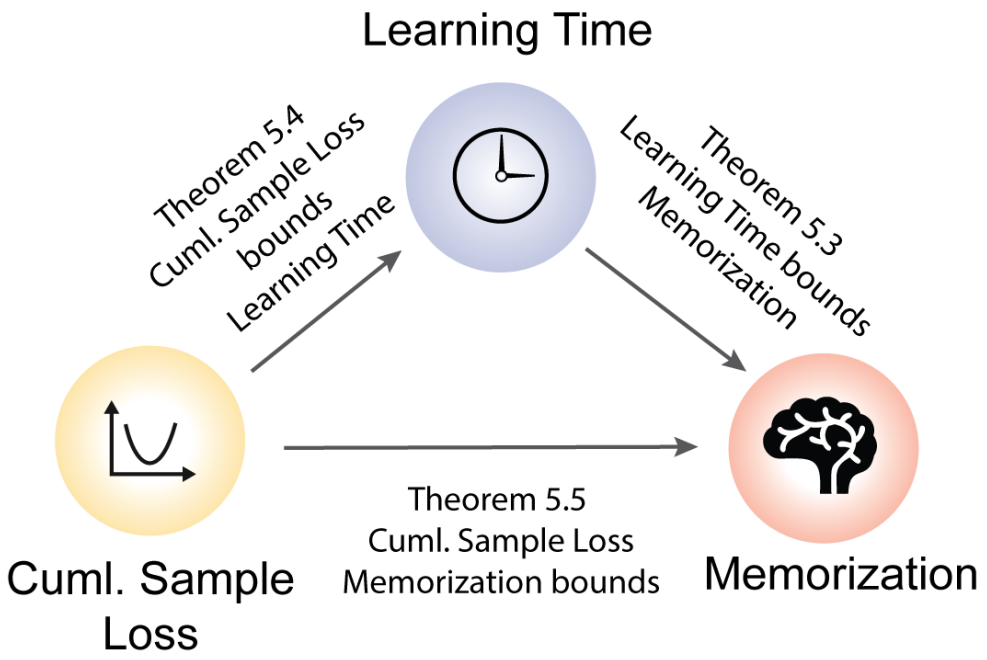
Towards Memorization Estimation: Fast, Formal and Free

Deepak Ravikumar[1], Efstathia Soufleri[1], Abolfazl Heshemi[1], Kaushik Roy[1]
[1]School of Electrical and Computer Engineering, Purdue University

ICML 2025
International Conference On Machine Learning

SRC

Overview



The Challenge: Understanding Memorization in Deep Learning

Our Contribution: Cumulative Sample Loss (CSL)

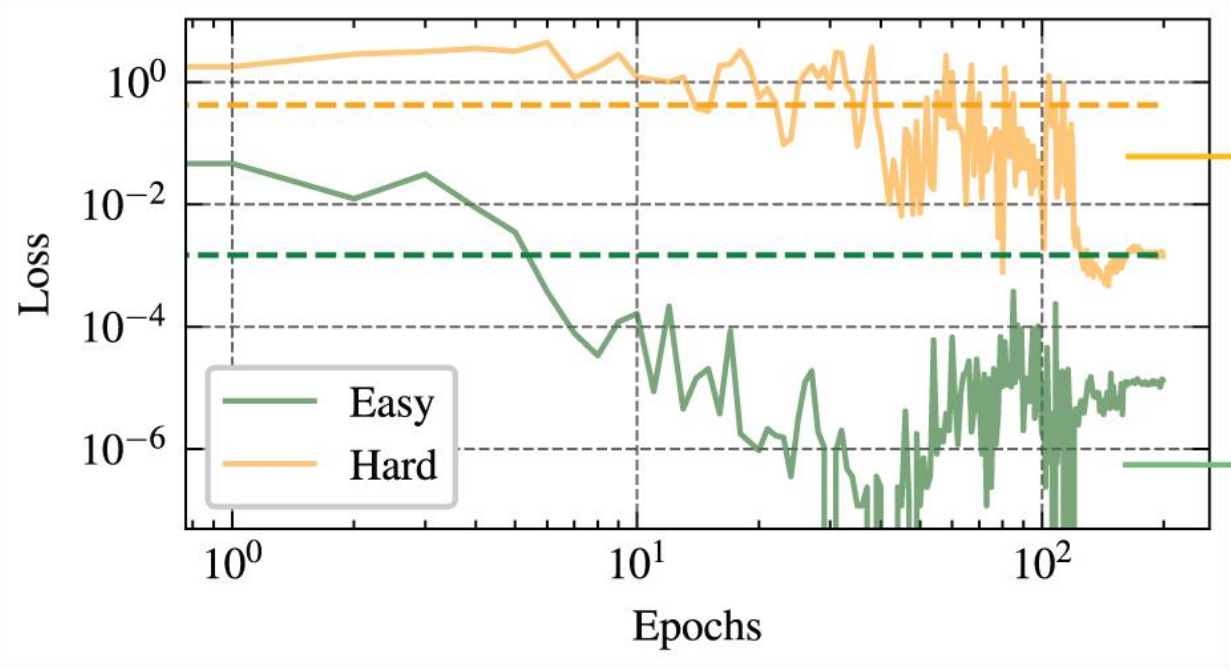
- We introduce a new, efficient proxy for memorization: Cumulative Sample Loss (CSL). CSL is simply the accumulated loss of a sample over the entire training process.
- We establish a theoretical framework that connects CSL to both learning time and stability-based memorization.

Key Benefits & Applications

- Fast & Free: CSL is 10,000x faster than stability-based methods and can be obtained with zero extra overhead during training.
- Practical Applications: State-of-the-art performance for identifying mislabeled examples and detecting duplicates in datasets.

SRC

Intuition and CSL



Learnt Late
Hard,
Atypical
mem. - 0.67

Learnt Early
Easy,
Typical
mem - 0.0

Visualizing peacock class learning in ImageNet. Average loss is shown for easy and hard-to-learn peacocks. The dashed line represents average loss, while solid lines show actual loss. Easy images are less memorized, while hard images are memorized more.

Define Learning Time and CSL

Cumulative Sample Loss: $CSL(\vec{z}_i) = \sum_{t=0}^{T-1} l(\vec{w}_t, \vec{z}_i)$

Sample Learning Condition: $\mathrm{E}\left[||\nabla_{x_i} l(\vec{w}_R)||_2^2\right] \leq \tau$

Learning Time: $T_{z_i} = min_T \left\{T: \mathrm{E}\left[||\nabla_{x_i} l(\vec{w}_R)||_2^2\right] \leq \tau\right\}$

SRC

Theoretical Results

Assumptions:

Bounded loss (or cross entropy), Lipschitznes, bounded gradient variance and unbiased grad. est.

Lemma 5.1: Input gradient norm is bound by weight gradient norm.

Theorem 5.2: Convergence in input gradient norm for SGD converges is root of iterations.

Theorem 5.3: Learning Time bounds Memorization.

$$\mathrm{E}_{z_i}[mem(\vec{z}_i)] \leq \kappa_T \, \mathrm{E}_{z_i}[T_{z_i}] + \frac{\beta}{L}$$

Theorem 5.4: Cumulative loss bounds learning time.

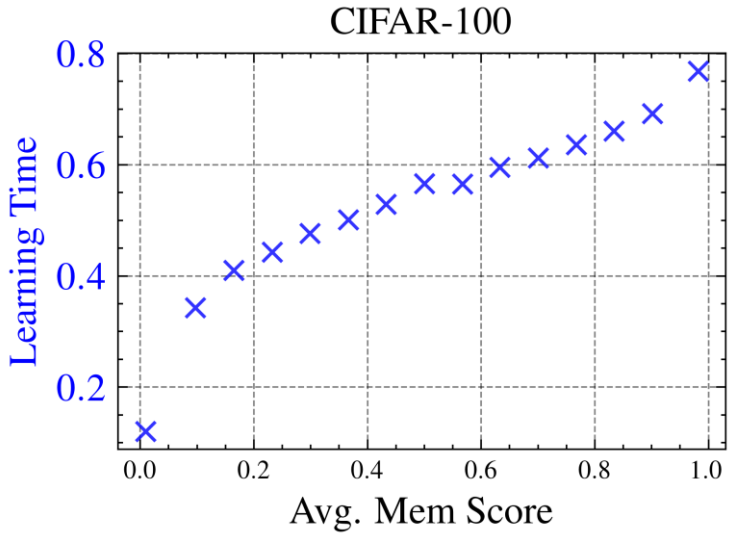
$$\kappa_T \, \mathrm{E}_{z_i}[T_{z_i}] \leq \frac{\mathrm{E}_{z_i}[CSL(\vec{z}_i)] - \xi}{L}$$

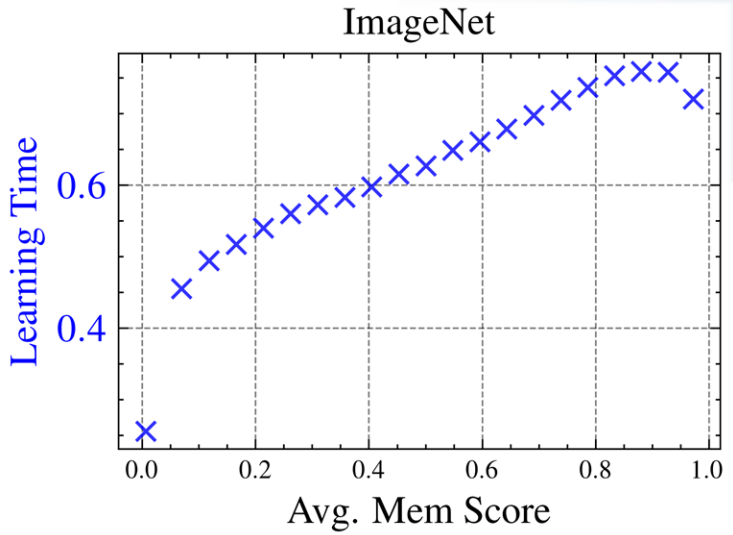
Theorem 5.5 Cumulative Sample Loss bounds Memorization.

$$\mathrm{E}_{z_i}[mem(\vec{z}_i)] \leq \frac{\mathrm{E}_{z_i}[CSL(\vec{z}_i)] + \beta - \xi}{L}$$

SRC

Validating Theoretical Results



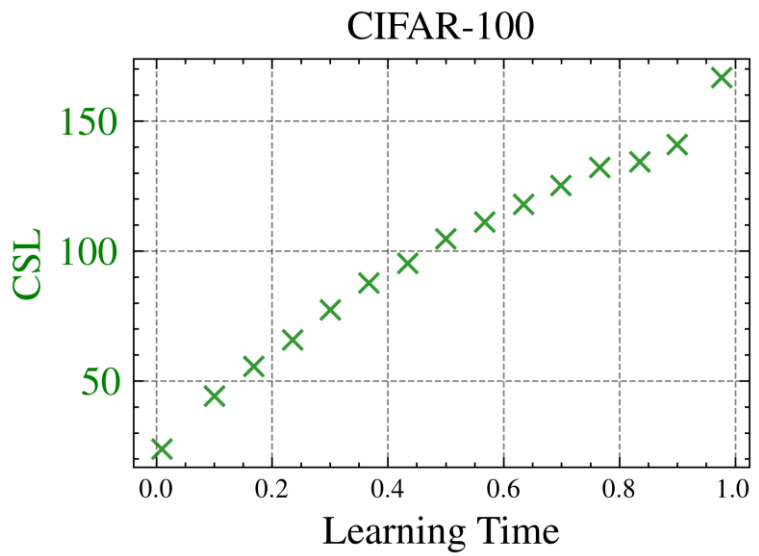


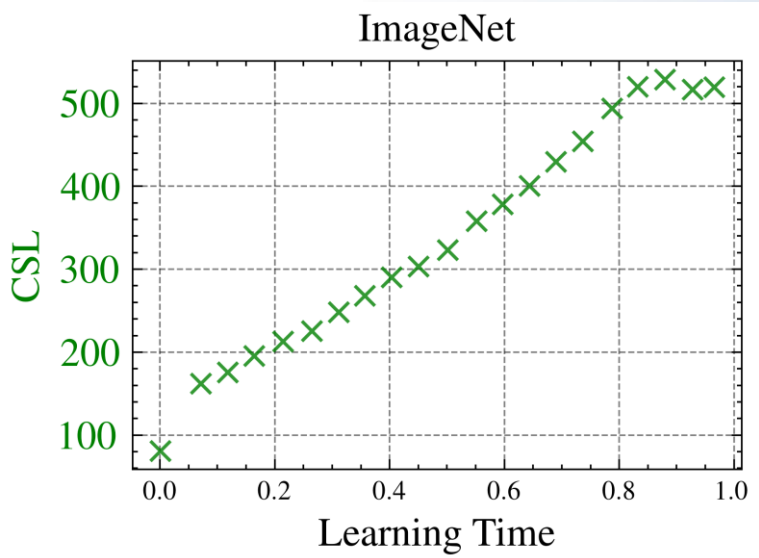
Theorem 5.3: Learning Time bounds Memorization.

$$\mathrm{E}_{z_i}[mem(\vec{z}_i)] \leq \kappa_T \, \mathrm{E}_{z_i}[T_{z_i}] + \frac{\beta}{L}$$

The expectation is that it has a linear relation between learning time and memorization score.
This is validated by results on CIFAR100 and ImageNet

Validating Theoretical Results



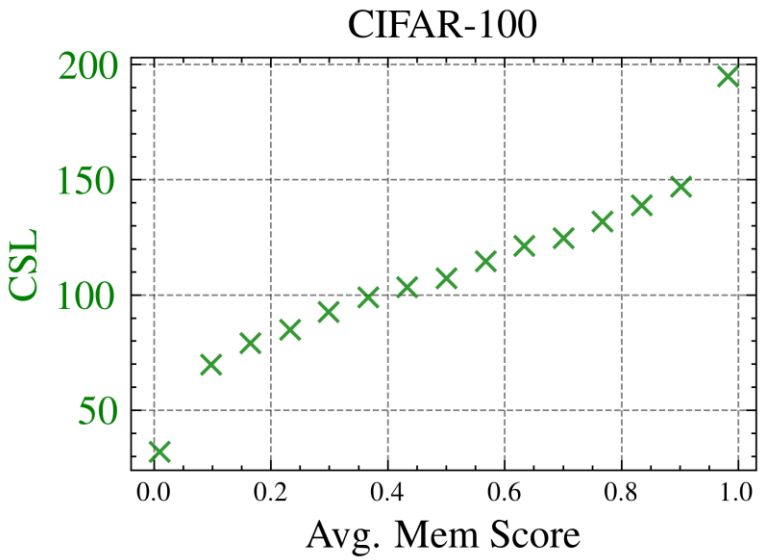


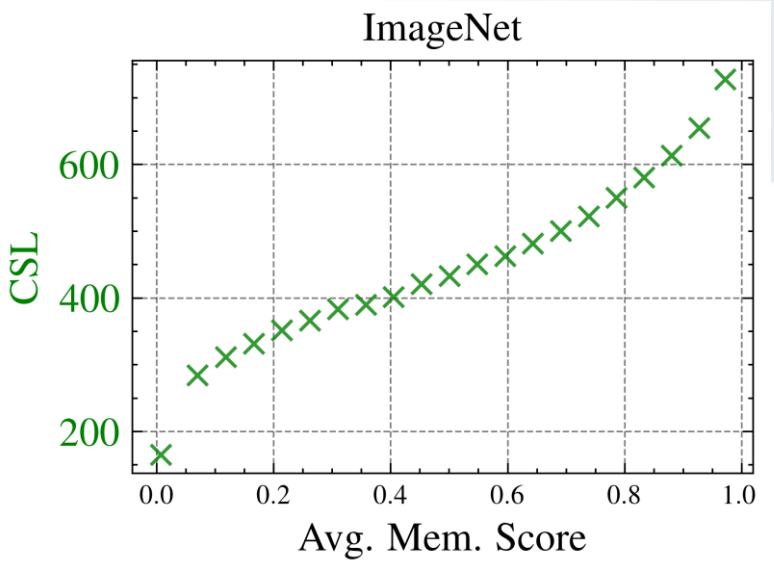
Theorem 5.4: Cumulative loss bounds learning time.

$$\kappa_T \, \mathrm{E}_{z_i}[T_{z_i}] \leq \frac{\mathrm{E}_{z_i}[CSL(\vec{z}_i)] - \xi}{L}$$

The expectation is that it has a linear relation between learning time and CSL. This is validated by results on CIFAR100 and ImageNet

Validating Theoretical Results





Theorem 5.5 Cumulative Sample Loss bounds Memorization.

$$\mathrm{E}_{z_i}[mem(\vec{z}_i)] \leq \frac{\mathrm{E}_{z_i}[CSL(\vec{z}_i)] + \beta - \xi}{L}$$

The expectation is that it has a linear relation between memorization and CSL. This is validated by results on CIFAR100 and ImageNet

Similarity with Memorization

Dataset	Arch.	Subset	Method	CS	PC
CIFAR-100	Inception	Top 5k	Final Sample Loss	0.33	0.06
			Curv	0.87	0.16
			Loss Sensitivity	**0.97**	0.39
			Forget Freq.	0.96	0.29
			CSL (Ours)	0.93	**0.4**
		All	Final Sample Loss	0.24	0.17
			Curv	0.69	0.49
			Loss Sensitivity	0.81	0.76
			Forget Freq.	0.76	0.59
			CSL (Ours)	**0.87**	**0.79**
ImageNet	ResNet50	Top 50k	Final Sample Loss	0.78	0.12
			Curv	0.84	0.05
			Loss Sensitivity	0.79	0.04
			Forget Freq.	0.68	0.15
			CSL (Ours)	**0.94**	**0.21**
		All	Final Sample Loss	0.64	0.5
			Curv	0.62	0.33
			Loss Sensitivity	0.49	0.17
			Forget Freq.	0.49	0.04
			CSL (Ours)	**0.79**	**0.64**

CSL correlation and similarity with memorization compared to other methods across CIFAR-100 and ImageNet datasets. CS denotes cosine similarity and PC denotes Pearson correlation.

Identifying Mislabeled

Dataset	Method	1% Noise	2% Noise	5% Noise	10% Noise
CIFAR-10	Thr. Learning Time	0.4951 ± 0.0248	0.4954 ± 0.0044	0.4911 ± 0.0071	0.4948 ± 0.0057
	In Conf.	0.8781 ± 0.0177	0.8072 ± 0.0130	0.7254 ± 0.0214	0.6528 ± 0.0042
	CL	0.8651 ± 0.0127	0.8905 ± 0.0115	0.8874 ± 0.0019	0.8551 ± 0.0030
	SSFT	0.9626 ± 0.0018	0.9551 ± 0.0020	0.9498 ± 0.0042	0.9360 ± 0.0020
	Curv.	0.9715 ± 0.0045	0.9776 ± 0.0033	0.9800 ± 0.0003	0.9819 ± 0.0006
	CSL (Ours)	**0.9845 ± 0.0026**	**0.9864 ± 0.0004**	**0.9870 ± 0.0003**	**0.9869 ± 0.0005**
CIFAR-100	Thr. Learning Time	0.5256 ± 0.0012	0.5227 ± 0.0100	0.5161 ± 0.0051	0.5203 ± 0.0029
	In Conf.	0.7258 ± 0.0102	0.7236 ± 0.0047	0.7069 ± 0.0069	0.6884 ± 0.0053
	CL	0.8723 ± 0.0208	0.8838 ± 0.0006	0.8733 ± 0.0010	0.8536 ± 0.0006
	SSFT	0.8915 ± 0.0045	0.8893 ± 0.0013	0.8784 ± 0.0030	0.8664 ± 0.0024
	Curv.	0.9856 ± 0.0009	0.9865 ± 0.0011	0.9876 ± 0.0021	0.9892 ± 0.0012
	CSL (Ours)	**0.9891 ± 0.0003**	**0.9895 ± 0.0002**	**0.9895 ± 0.0001**	**0.9897 ± 0.0001**

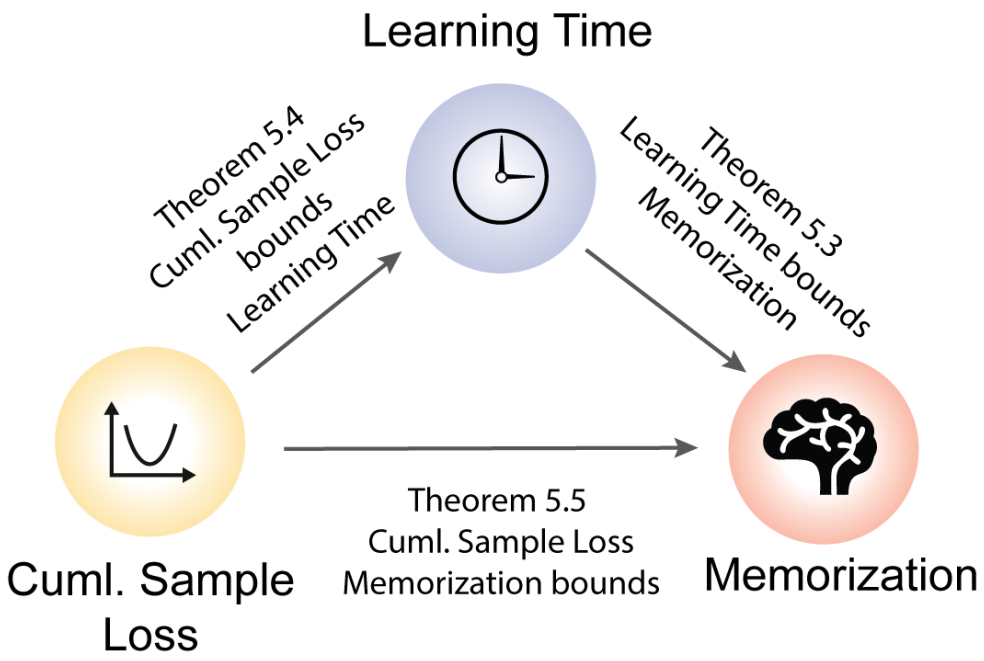
Evaluating the performance of mislabeled detection of the proposed framework against existing methods on CIFAR-10 and CIFAR-100 datasets under various levels of label noise.

Identifying Duplicate Samples

Method	CIFAR-10	CIFAR-100
Thr. LT	0.7029 ± 0.0058	0.7419 ± 0.0059
In Conf.	0.9237 ± 0.0114	0.8623 ± 0.0131
CL	0.5533 ± 0.0031	0.5873 ± 0.0090
SSFT	0.8490 ± 0.0034	0.7938 ± 0.0045
Curv.	0.9536 ± 0.0030	0.9639 ± 0.0030
CSL (Ours)	**0.9821 ± 0.0006**	**0.9886 ± 0.0008**

Result of duplicate detection using the proposed methods and other baselines on CIFAR-10 and CIFAR-100 datasets.

Summary



The Challenge: Understanding Memorization in Deep Learning

Our Contribution: Cumulative Sample Loss (CSL) with thiery

Benefits & Applications

- Fast & Free: CSL is 10,000x faster than stability-based methods and can be obtained with zero extra overhead during training.
- Practical Applications: State-of-the-art performance for identifying mislabeled examples and detecting duplicates in datasets.

Thank you!