On Information Captured by Neural Networks

Connections with Memorization and Generalization

Hrayr Harutyunyan

Rising Stars in AI Symposium 2023

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Information Sciences Institute

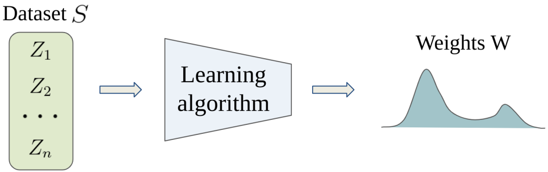
USC Viterbi
School of Engineering

Works discussed in this talk

- H, Reing, Ver Steeg, Galstyan. **Improving generalization by controlling label-noise information in neural network weights.** ICML 2020.
- H, Achille, Paolini, Majumder, Ravichandran, Bhotika, Soatto. **Estimating informativeness of samples with smooth unique information.** ICLR 2021.
- H, Raginsky, Ver Steeg, Galstyan. **Information-theoretic generalization bounds for black-box learning algorithms.** NeurIPS 2021.
- H, Ver Steeg, Galstyan. **Formal limitations of sample-wise information-theoretic generalization bounds.** IEEE ITW 2022.

Why and how do neural networks generalize?

Information-theoretic perspective

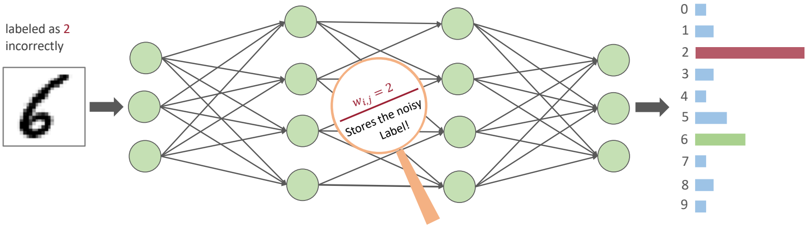


- How to measure information?
- What kind of information should we measure?
- How to quantify memorization?
- How to reduce some forms of memorization?
- How is information captured by neural networks related to generalization?

Learning setting

1. Input space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{1, 2, \ldots, C\}$.
2. Training set $S = (Z_1, \ldots, Z_n)$ consisting of n i.i.d. samples from a distribution P_Z on $\mathcal{Z}$.
    - $\boldsymbol{X} \triangleq (X_1, \ldots, X_n)$, $\boldsymbol{Y} \triangleq (Y_1, \ldots, Y_n)$.
3. Hypothesis space $\mathcal{W}$.
4. Training algorithm $Q_{W|S}$ (a probability kernel), which takes a training set and returns a distribution on hypotheses.
5. Loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$.
6. Empirical risk: $r_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i)$.
7. Population risk: $R(w) = \mathbb{E}_{Z' \sim P_Z} \left[\ell(w, Z')\right]$.

Label-noise memorization



Label-noise information can be measured by $I(W; \boldsymbol{Y} \mid \boldsymbol{X})$.

- ERM with cross-entropy loss maximizes label-noise information.
- Small $I(W; \boldsymbol{Y} \mid \boldsymbol{X})$ implies prediction "mistakes" on incorrectly labeled examples.
- Minimizing $I(W; \boldsymbol{Y} \mid \boldsymbol{X})$ improves a generalization gap bound.

Label-noise memorization

The proposed method for limiting label-noise information

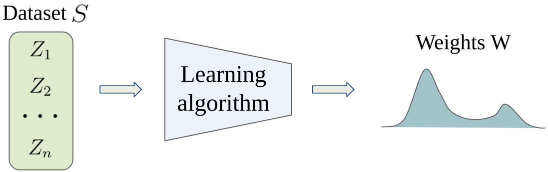
We derive a training algorithm that minimizes empirical risk subject to limited label-noise information $I(W; \boldsymbol{Y} \mid \boldsymbol{X})$.

Method	no noise	uniform noise				pair noise			
	0%	20%	40%	60%	80%	10%	20%	30%	40%
ERM with cross entropy loss	92.7	85.2	81.0	69.0	38.8	90.0	88.1	87.2	81.8
Proposed	93.3	**92.2**	**90.2**	**82.9**	**44.3**	**93.0**	**92.3**	**91.1**	**90.0**

Table 1: Test accuracy comparison on CIFAR-10, corrupted with various label noise types.

A more general notion of memorization

How much information does a particular example provide to the training of a neural network?



High-level summary of our work

We propose to consider $I(W; Z_i = z_i \mid Z_{-i} = z_{-i})$ or its function space analog $I(\widehat{Y}; Z_i = z_i \mid Z_{-i} = z_{-i}, X = x)$ as a measure of memorization/informativeness.

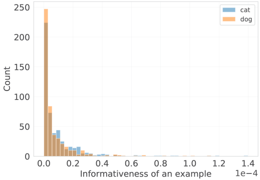
- Not necessarily harmful memorization.
- Relates to the question "what will happen if remove the example?".

Which examples are most informative?

(a) Least informative examples

(b) Most informative examples

(c) Histogram of informativeness scores



Main findings

- Most examples have small information content.
- Outliers, hard examples, and rare examples are more informative.
- Examples with incorrect labels are informative (as their label is memorized).
- Different networks agree well on which examples are informative.
- Examples of challenging datasets are more informative on average.

Information-theoretic generalization bounds

Theorem (Xu & Raginsky [1]; Bu, Zou, Veeravalli [2])

Let $W \sim Q_{W|S}$. *If* $\ell(w, z) \in [0, 1]$ *then*

$$\begin{aligned}\underbrace{\left|\mathbb{E}_{S,W}\left[R(W) - r_S(W)\right]\right|}_{\textit{exp. generalization gap}} &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2}I(W; Z_i)} \\ &\leq \sqrt{\frac{1}{2n}I(W; S)} \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2}I(W; Z_i \mid Z_{-i})}.\end{aligned}$$

---

[2] Xu and Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. NeurIPS 2017.

[2] Bu, Zou, Veeravalli. Tightening mutual information-based bounds on generalization error. IEEE JSAIT 2022

High-level summary of our contribution

Our main contribution

We derive generalization bounds based on the information contained in predictions rather than weights.
The core idea is to encode the learned function with a random variable.

A general learning algorithm setting:

- The learning algorithm $f : \mathcal{Z}^n \times \mathcal{X} \times \mathscr{E} \to \widehat{\mathcal{Y}}$ that takes a training set z, a test input x', an auxiliary argument ε capturing any stochasticity, and outputs a prediction $f(z, x', \varepsilon)$ on the test example.

- $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ measures the discrepancy between a prediction and a label.
- Empirical risk: $r_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(S, X_i, \mathcal{E}), Y_i)$.
- Population risk: $R(f) = \mathbb{E}_{Z' \sim P_Z} \left[\ell(f(S, X', \mathcal{E}), Y')\right]$.

The setting of Steinke and Zakynthinou (2020)[3]

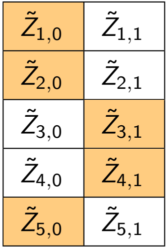
- Let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ be a collection of $2n$ i.i.d. samples from P, grouped into n pairs.
- $J \sim \text{Uniform}(\{0,1\}^n)$ specifies which example to select from each pair to form the training set:

$$S = (\tilde{Z}_{i,J_i})_{i=1}^{n}.$$

Example 1

$J = (0,0,1,1,0)$

$\tilde{Z}_J$



$\tilde{Z}_{1,0}$	$\tilde{Z}_{1,1}$
$\tilde{Z}_{2,0}$	$\tilde{Z}_{2,1}$
$\tilde{Z}_{3,0}$	$\tilde{Z}_{3,1}$
$\tilde{Z}_{4,0}$	$\tilde{Z}_{4,1}$
$\tilde{Z}_{5,0}$	$\tilde{Z}_{5,1}$

---

[3] Steinke and Zakynthinou. Reasoning about generalization via conditional mutual information. COLT 2020.

Functional CMI generalization gap bound

Theorem

If $\ell(\widehat{y}, y) \in [0,1], \forall \widehat{y} \in \widehat{\mathcal{Y}}, y \in \mathcal{Y}$, *then*

$$\underbrace{\left|\mathbb{E}_{\tilde{Z},J,\mathcal{E}}\left[R(f) - r_S(f)\right]\right|}_{\text{exp. generalization gap}} \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\tilde{z}\sim\tilde{Z}}\sqrt{2I(\underset{\uparrow}{f(\tilde{z}_J, \tilde{x}_i, \mathcal{E})}\,;\,\underset{\uparrow}{J_i})}.$$

predictions on the i-th pair

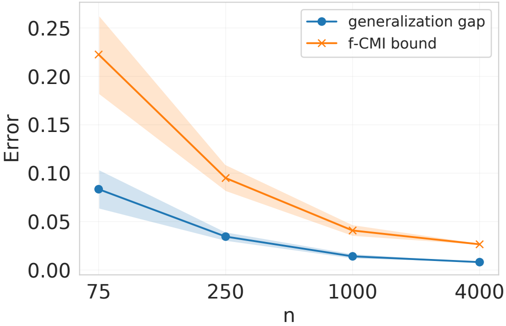
train-test split
variable of the i-th pair

Benefits:

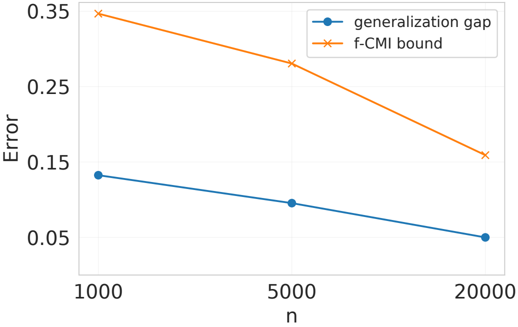
- The right-hand side depends on MIs between low-dimensional variables.
- Finite VC dimensionality d implies an $\tilde{O}(\sqrt{d/n})$ information-theoretic bound.
- On-average stability implies a small information-theoretic bound.

Experimental Results

Setup: MNIST 4 vs 9 classification with 4-layer CNN (3M parameters, deterministic algorithm).



Setup: Fine-tuning a pretrained ResNet-50 on CIFAR-10 (SGD with momentum + data augmentations).



Expected vs *expected squared* generalization gap bounds

Expected generalization gap bounds:

$$\left|\mathbb{E}_{S,W}\left[R(W) - r_S(W)\right]\right| \leq \underbrace{\frac{c}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)}}_{\text{sample-wise bound}} \leq \underbrace{c\sqrt{\frac{I(W;S)}{n}}}_{\text{whole dataset information bound}} .$$

Expected *squared* generalization gap bounds:[4,5]

$$\mathbb{E}_{W,S}\left[\left(R(W) - r_S(W)\right)^2\right] \leq \text{a sample-wise bound?} \leq \underbrace{\frac{I(W;S)+c}{n}}_{\text{whole dataset information bound}} .$$

---

[5] Harutyunyan, Raginsky, Ver Steeg, Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. NeurIPS 2021.

[5] Aminian, Toni, Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. IEEE ISIT 2021.

A limitation of sample-wise information measures

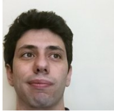
Main results

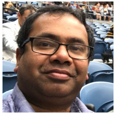
1. Sample-wise expected squared, PAC-Bayes, and single draw generalization bounds do not exist.
2. Starting at subsets of size 2, there are expected squared generalization gap bounds that measure information between W and a subset of examples.

$$\mathbb{E}_{S,W}\left[\left(R(W)-r_S(W)\right)^2\right] \leq \frac{1}{n} + \frac{1}{n^2}\sum_{i\neq k}\sqrt{2I\left(W;Z_i,Z_k\right)}.$$

3. These results hold for more advanced sample-wise bounds as well.

Thank you


Alessandro Achille


Rahul Bhotika


Orchid Majumder

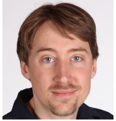

Giovanni Paolini


Maxim Raginsky


Avinash Ravichandran

Kyle Reing


Stefano Soatto


Greg Ver Steeg


Aram Galstyan

Find me at `hrayrhar.github.io`