Learning and Memorization

Bernhard Gstrein

Motivation

- [Zhang et al. 2017]: neural networks have the capacity to memorize their training set
  - Train AlexNet on CIFAR-10 with randomly permuted labels
  - Training error goes to 0
- What is the link between **memorization** and **generalization**?
  - Why don't NNs just memorize their training set?
- [Chatterjee 2018]: Is it possible to generalize by memorizing alone?

Basic idea of paper

- What is a simple form of memorization? $\rightarrow$ a table

Lives in water	Has eyes	Has limbs	Vertebrate
0	1	1	1
1	1	0	1
1	0	0	0

Model for classification of animals into vertebrates/invertebrates

- We must binarize the dataset
- We must limit the complexity
  - 28x28 images $\rightarrow$ $28 \cdot 28 = 784$ $\rightarrow$ $2^{784} \propto 10^{236}$

Table of Contents

Preprocessing data

- MNIST dataset: 28x28 images of handwritten digits (0-9)
- We unroll the images: $28 \cdot 28 = 784$
- We scale the numerical values to the range $[0, 1]$
- We binarize the data using the operator > 0.5
- Labels (to be predicted): class 0-4 vs. 5-9

- We end up with
  - Features: matrix of shape $(N, 784)$, boolean entries
  - Labels: matrix of shape $(N, 1)$, boolean entries

A single lut

- Reminder: every example is an instance of a "bit pattern" (e.g. $\boldsymbol{x}^i = 010$) and has a label (e.g. $y^i = 1$)
- For each bit pattern, we cound how many times $y = 0$ and $y = 1$

$$\hat{f}(\text{bit pattern}) = \begin{cases} 0 & \text{if} \quad \sum\limits_{y=0} > \sum\limits_{y=1} \\ 1 & \text{if} \quad \sum\limits_{y=0} < \sum\limits_{y=1} \\ \text{rand}(0,1) & \text{if} \quad \sum\limits_{y=0} = \sum\limits_{y=1} \end{cases}$$

A single lut: example

$\boldsymbol{x}$	y
000	0
000	1
000	1
001	1
100	0
110	0
110	1

bit pattern	$\sum_{y=0}$	$\sum_{y=1}$
000	1	2
001	0	1
010	0	0
011	0	0
100	1	0
101	0	0
110	1	1
111	0	0

bit pattern	$\hat{f}$
000	1
001	1
010	0*
011	1*
100	0
101	1*
110	1*
111	0*

A single lut applied on MNIST

- MNIST features: matrix of shape $(N, 784)$
- We perform PCA and obtain a matrix of shape (N, k), varying k from 2 to 20
- A single lut is able to handle this dataset

A single lut applied on MNIST

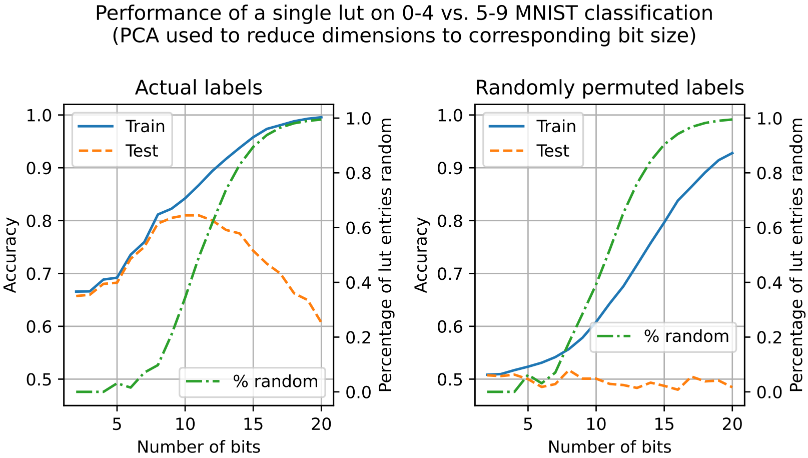

Table of Contents

Network

- As we've seen, a single lut is not very powerful

Table of Contents

How to go from there

Table of Contents

Recap

Hello there, this is empty :)

Any Questions?

Table of Contents

References lala

Chatterjee, Satrajit (2018). “Learning and memorization”. In: *International Conference on Machine Learning*. PMLR, pp. 755–763.

Zhang, Chiyuan et al. (2017). *Understanding deep learning requires rethinking generalization*. arXiv: 1611.03530 [cs.LG].