PRIN17 - Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond

Kick off meeting - 14.10.19

Dipartimento Informatica - Unità UNIMI

Componenti Unità

Marco Frasca	RTD A (responsabile)
Giorgio Valentini	PO
Dario Malchiodi	PA
Marco Mesiti	PA
Paolo Perlasca	RU

UNIVERSITÀ DEGLI STUDI DI MILANO

PRIN17 - UNIMI Research Activity

[T1 - D1.1] Compare classic Data Structures vs Purely Learned Indexes - Characterization of the space-time-accuracy performance of ML models in terms of distribution of the input data, and their comparison against the known (compressed) data structures

Collaboration with UNIPA unit

- Preliminary experiments about learned indexes with neural networks (NN) models [UNIPA] + NN compression techniques [UNIMI] to reduce the space occupancy + their comparison with some SOA indexes [UNIPA]
- PROBLEM INPUT: sorted sequence of integers X [extendable to other type of elements]
- PROBLEM OUTPUT: a learned indexed for searching elements in X
- STEP 1: Coding the input for feeding the NN [e.g. binary representation]
- STEP 2: Coding the labels for training $(x \in X, \hat{F}(x))$
- STEP 3: NN training
- STEP 4 [optional]: NN compression
- STEP 5: Index validation

PRIN17 - UNIMI Research Activity

[T1 - D1.1]. Some preliminary results.
X uniformly distributed

NN model	Compression rate	max error	max error/size	mean error/size
		Input size $1.05 \cdot 10^6$		
NN3	4.73	1270|660	$1.2 \cdot 10^{-1} \| 6.3 \cdot 10^{-2}$	$2.40 \cdot 10^{-4} \| 1.45 \cdot 10^{-4}$
NN2	4.39	1031|699	$9.0 \cdot 10^{-2} \| 6.7 \cdot 10^{-2}$	$2.33 \cdot 10^{-4} \| 1.45 \cdot 10^{-4}$
		Input size 8192		
NN3	4.75	40|32	$4.8 \cdot 10^{-1} \| 3.9 \cdot 10^{-1}$	$1.14 \cdot 10^{-3} \| 1.24 \cdot 10^{-4}$
NN2	4.4	33|34	$4.0 \cdot 10^{-1} \| 4.1 \cdot 10^{-1}$	$1.12 \cdot 10^{-3} \| 1.19 \cdot 10^{-3}$

Ongoing activities:
Experiments with different data distributions

Extending the learned index in the sense of Kraska *et al.* 2017 (also related to Task 3)

PRIN17 - UNIMI Research Activity

[T1 - D1.1]. Short- Mid-term objectives

- Characterize the effectiveness of NN compression techniques with different data distributions
- Actually we can relate the expected space occupancy to the compression rate or designing the model to reduce the *mean* error. We aim at investigating whether it is possible to relate (empirically at least) space occupancy with *maximum* error
- Extending the learned index to more levels and evaluate how the compression effectiveness (space - accuracy) varies with it
- Comparing with other 'classical' indexes

PRIN17 - UNIMI Research Activity

[T2 - D2.1] Compress ML models - Development of compressed ML models, by investigating the trade-off between compressed space, compressed time and the prediction accuracy

Just UNIMI unit (so far ...)

- Preliminary experiments to evaluate the capability of different compression techniques for NN models to reduce the model space requirements while preserving (or improving) its accuracy
  - Pruning
  - Weight sharing
  - Sparsified NN training
  - ...
  - Combined approaches
- Results might change with the input problem
- NN models for different categories of problems: multi-class/binary classification, regression, dimensionality reduction and so forth

[T2 - D2.1]. Some preliminary results

MNIST digit classification

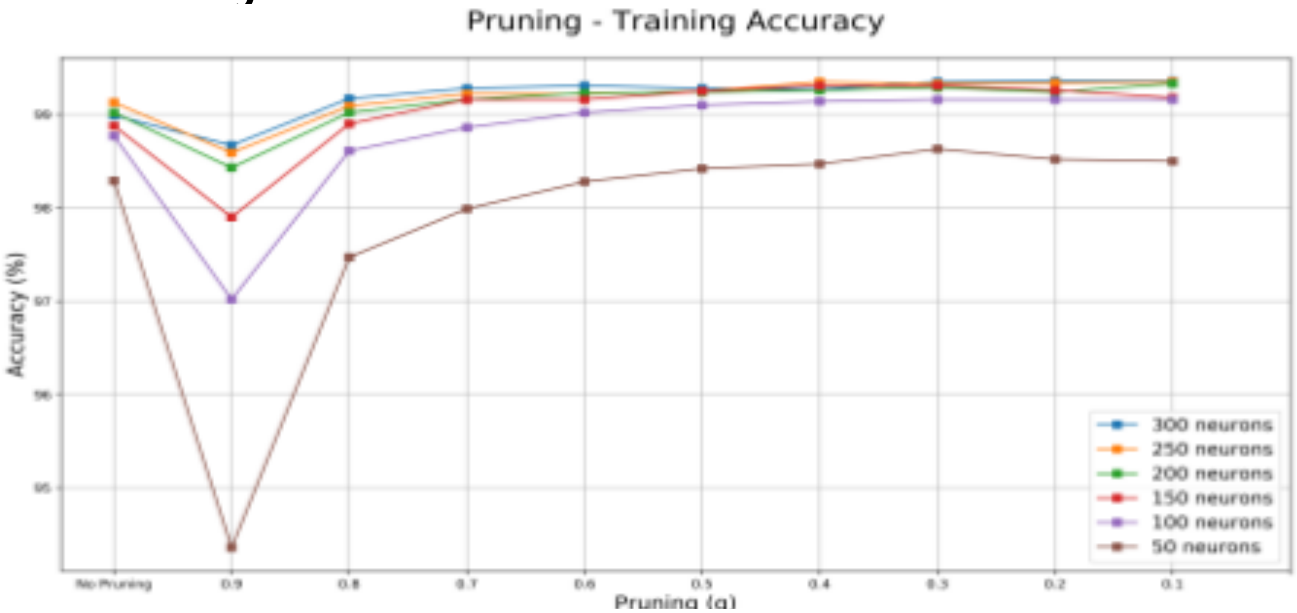
INPUT: feature vector (dim 784) representing images of hand-written digits

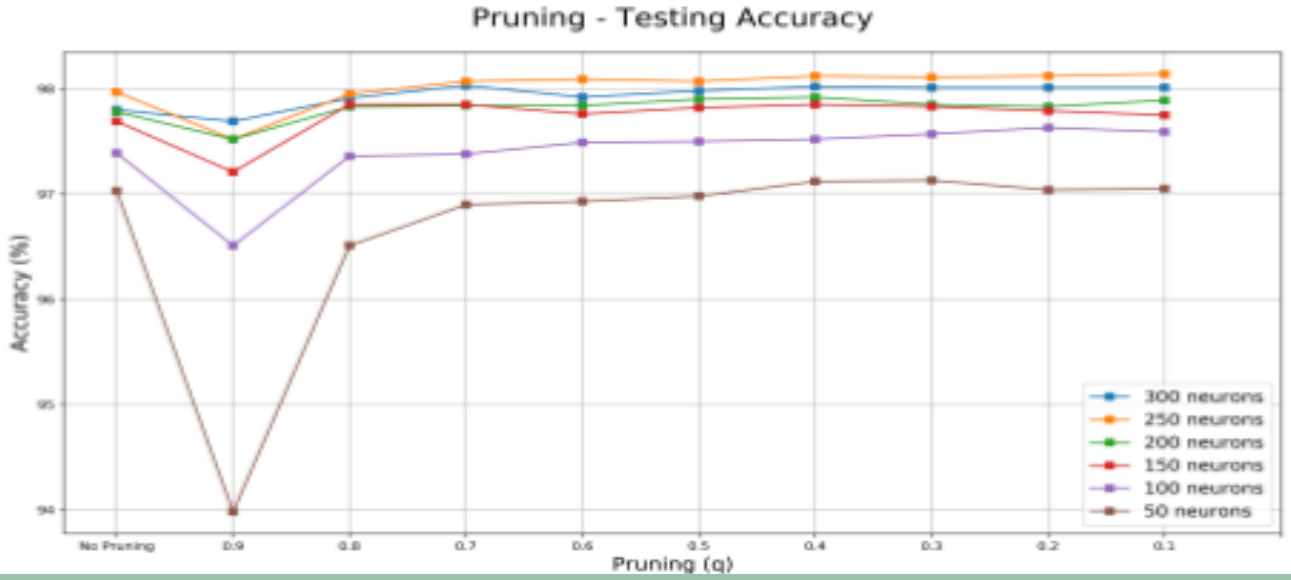
OUTPUT: softmax 10-neurons layer

Model structure: FFW with 1 hidden layer (variable size)

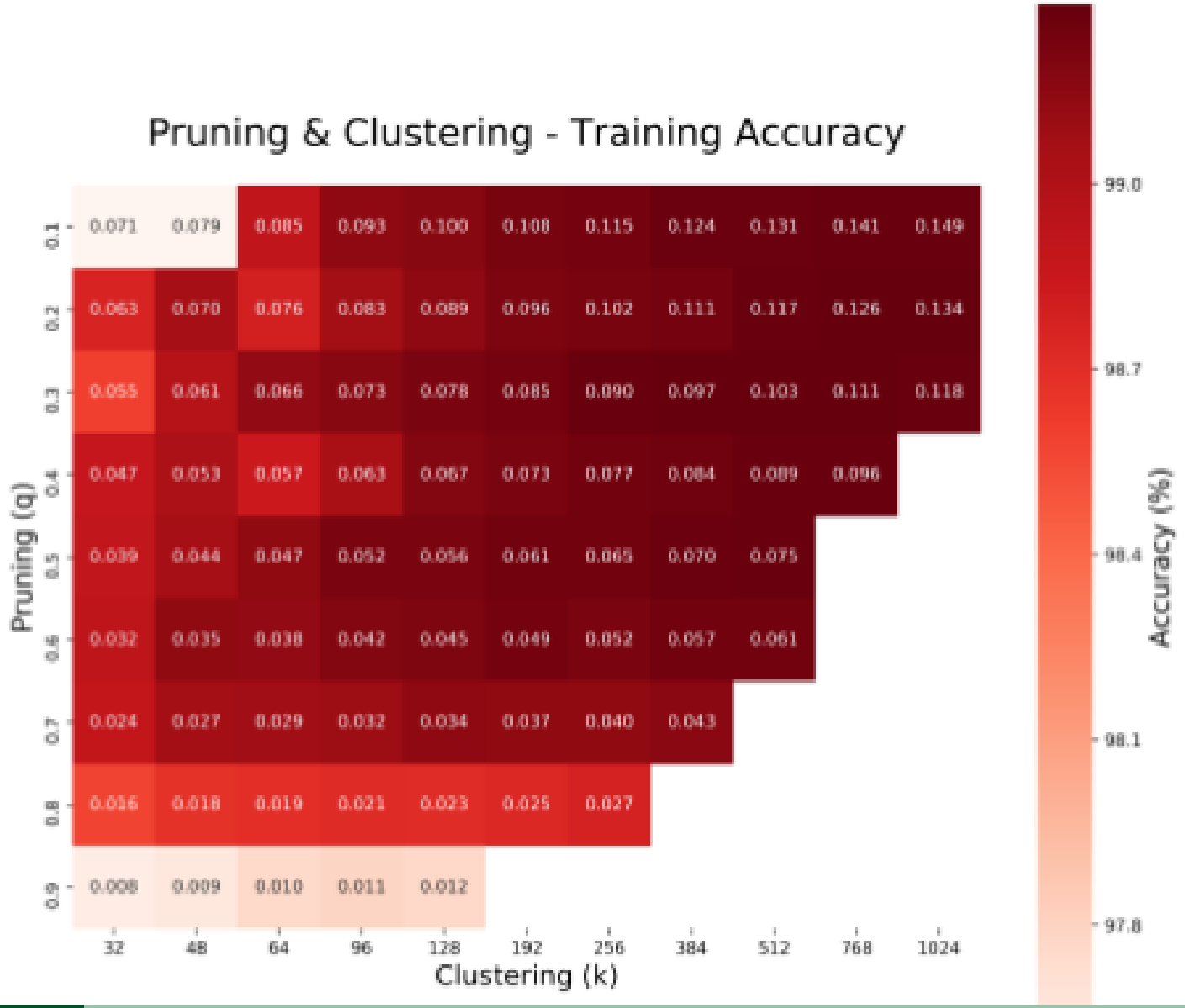
Ongoing activities: Studying novel techniques for NN compression

Experiments with different problem categories





[T2 - D2.1]. Some preliminary results



PRIN17 - UNIMI Research Activity

[T2 - D2.1]. Short- Mid-term objectives

- Applying different NN models
- Exploring/designing other NN compression methods. In particular, merging the NN learning and compression phases
  - With the aim of defining a multi-criteria training algorithms (optimizing space and accuracy)
- Compressing other ML models?

PRIN17 - UNIMI Research Activity

We can also give a contribution to

[T1 - D1.2] A collection of known and new implementations of ML-based and compressed data structures, to be used in the next tasks T3 and T4 as "building blocks" for our multicriteria framework

[T2 - D2.2] A collection of software prototypes for succinct ML-models

Possible other colloborations are welcome