Relation of work project for Intelligent Systems course

Compressed Network in EML

Ugo Marchesini 1

|  |
| --- |
| **Citation:** Lastname, F.; Lastname, F.; Lastname, F. Title. *Appl. Sci.* **2022**, *12*, x. https://doi.org/10.3390/xxxxx  Academic Editor: Firstname Lastname  Received: date  Accepted: date  Published: date  **Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.    **Copyright:** © 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Affiliation 1; ugo.marchesini@studio.unibo.it

**Abstract:** le reti neurali sono delle applicazioni del Machine Learning , permettono di risolvere problemi di regressione e classificazione modificando i propri pesi e bias durante la fase di addestramento. Reti più compresse in generale premettono di risolvere problemi con dataset più complessi , ma a costo dell’aumento del footprint della rete in termini di tempo necessario per l’addestramento , la predizione , la dimensione su disco e il carico della rete in caso di applicazione eseguita in remoto. Anche reti neurali con precisione molto elevata non forniscono una chiara comprensione della natura del problema né del modo in cui la rete “risolve” il problema.Una rete neurale modifica i propri pesi e bias per meglio adattare il valore predetto con il valore atteso al fine di migliorare la precisione della predizione, è chiaro che all’aumentare della dimensione dei parametri addestrabili aumenti il costo in tempo ed in energia dell’addestramento stesso. Reti come le CNN (ResNet,Inception,VGG) contengono fino a decine di milioni di pesi , la loro precisione può essere molto alta , ma è anche molto alto il loro footprint. Le tecniche di compressione delle reti neurali permettono di ottenere delle footprint ridotte con perdite accettabili di precisione. Questo non risolve i problemi di ottimizzazione , dove è necessario imporre una funzione obbiettivo e definire dei vincoli tra variabili.EML permette in incorporare un modello ML in un problema combinatorio , in questo caso una rete neurale. Un problema combinatorio essendo dichiarativo potrebbe permettere una migliore comprensione e quindi una migliore “spiegabilità” del problema appreso da una rete neurale. Il campo di applicazione di questa ricerca è la transprecision computing. Lo scopo di questa ricerca è l’esplorazione delle tecniche di compressione delle reti neurali e la sua incorporazione in un modello di ottimizzazione combinatoria al fine di ottenere una pipeline che processi un dataset ottenuto dall’elaborazione di diverse funzioni di calcolo su diversi hardware.

**Keywords:** rete neurale; machine learning; footprint;compression; ottimizzazione combinatoria;

transprecision computing

1. Transprecision computing

Transprecision Computing is a paradigm that allows users to trade the energy associated with computation in exchange for a reduction in the quality of the computation results. Guaranteed numerical precision of each elementary step in a complex computation has been the mainstay of traditional computing systems for many years. La possibilita di ottenere una precisione accettabile al di sotto di un errore predeterminato con un utilizzo minore di bit nelle operazione a FP permette un risparmio energetico apre alla possibilità di …

1. Empirical Model Learning

EML can be considered as a technique that allow to encapsulate the learned model in a number of optimization techniques.

1. Compressed Neural Network

Tra le varie tecniche di compressione la ricerca ha preso in considerazione : weight pruning e quantization , dove weight pruning si intende mettere a zero il valore del peso. E’ stata utilizzata il TensorFlow Model Optimization Toolkit per entrambe le tecniche

1. Dataset

Il dataset è stato preso da <https://zenodo.org/record/6575841#.YtE3kS8QNQJ> , con successive integrazioni.

Il dataset è composto da 9 file divisi per benchmark e hardware.

Ci sono 3 tipi di benchmark : Convolution , Correlation e Saxpy , e corrispondono ad altrettanti algoritmi utilizzati per produrre dei sample.

Ci sono 3 tipo di hardware : g100,pc e vm.

Ogni file ha un nome del tipo <benchmark>\_<hardware>.csv

I sample contengono delle variabili (var\_0 , var\_1 … var\_n) il cui contenuto è il numero di bit usati nel calcolo ,l’errore (error) , il tempo impiegato (time) , l’utilizzo medio della memoria (memory\_mean) e l’utilizzo della memoria di picco (memory\_peak).

1. Used Tools

L’ambiente utilizzato è Visual Studio 2022 , il linguaggio Python per la parte di ML e C# per la data preparation. Le librerie sono : TensorFlow\Keras , Emlib e Util.

1. Process

First Phase : Epidemics , BuildTrainPrintSave , Quantization , Pruning , Compress

Second Phase : BuildTrainPrintSave , Pruning

In una prima fase sono state esplorate le tecniche di compressione,quantizzazione e pruning su un problema epidemiologico , e successivamente sul problema di transprecision computing.

* 1. Epidemics

Il primo esempio su cui esplorare le tecniche di compressione è stato un dataset di un modello epidemico di tipo SIR (Susceptible, Infectious, or Recovered).

Lo scopo del modello è la predizione dei valori SIR e la funzione obbiettivo del problema di ottimizzazione combinatoria è la minimizzazione del costo di un intervento NPI (non pharmaceutil interventions)

* + 1. Build and Train Model

Sono state usate delle funzioni per la creazione dei dati di input e di output configurabili. Le funzioni creano i dati ex novo se richiesto e li salvano , oppure caricano i dati se esistono.

Sono state addestrate due reti, una con DecisionTreeRegressor e l’altra con modelli Keras con 3 topologie diverse. Sostanzialmente differiscono solo per la profondità e il numero di neuroni per layer.

Tulle le reti sono salvate con estensione h5.

* + 1. Evaluate

La valutazione delle reti è stata effettuata con l’accuracy ottenuta durante la la fase di training , la rmse (root mean square error), mae (mean absolute error) e r2 (coefficient of determination).

Il dataset di test è il 20% del dataset di train.

* + 1. Post Training Quantization

Secondo la documentazione di TensorFlow : “Post-training quantization is a conversion technique that can reduce model size while also improving CPU and hardware accelerator latency, with little degradation in model accuracy. You can quantize an already-trained float TensorFlow model when you convert it to TensorFlow Lite format using the TensorFlow Lite Converter.”

Sono offerte 6 tecniche di quantizzazione : Dynamic range quantization, Full integer quantization, Integer with float fallback, Float16 quantization, Integer only: 16-bit activations with 8-bit weights (V1) e Integer only: 16-bit activations with 8-bit weights + builtins (V2)

Al fine di valutare le metriche di queste tecniche è necessario effettuare il training del modello (baseline) prima della conversione ad un modello quantizzato e compresso in formato tflite.

Le metriche sono : size , rmse, mae e r2;

* + 1. Pruning
    2. Compress
    3. Pipeline
  1. Transprecision Computing
     1. Build and Train Model
     2. Evaluate
     3. Pruning

Il primo esempio su cui esplorare le tecniche di compressione è stato un dataset di un modello epidemico di tipo SIR.

1. Metrics
2. Result

3. Results

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

3.1. Subsection

3.1.1. Subsubsection

Bulleted lists look like this:

* First bullet;
* Second bullet;
* Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

3.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.



**Figure 1.** This is a figure. Schemes follow the same formatting.

**Table 1.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |
| --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** |
| entry 1 | data | data |
| entry 2 | data | data 1 |

1 Tables may have a footer.

The text continues here (Figure 2 and Table 2).

|  |  |
| --- | --- |
| C:\Users\martin\Downloads\testFigure.tif | C:\Users\martin\Downloads\testFigure.tif |
| (**a**) | (**b**) |

**Figure 2.** This is a figure. Schemes follow another format. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel; (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

**Table 2.** This is a table. Tables should be placed in the main text near to the first time they are cited.

|  |  |  |  |
| --- | --- | --- | --- |
| **Title 1** | **Title 2** | **Title 3** | **Title 4** |
| entry 1 \* | data | data | data |
| data | data | data |
| data | data | data |
| entry 2 | data | data | data |
| data | data | data |
| entry 3 | data | data | data |
| data | data | data |
| data | data | data |
| data | data | data |
| entry 4 | data | data | data |
| data | data | data |

\* Tables may have a footer.

3.3. Formatting of Mathematical Components

This is example 1 of an equation:

|  |  |
| --- | --- |
| a = 1, | (1) |

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is example 2 of an equation:

|  |  |
| --- | --- |
| a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z | (2) |

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

**Theorem 1.** Example text of a theorem. Theorems, propositions, lemmas, etc. should be numbered sequentially (i.e., Proposition 2 follows Theorem 1). Examples or Remarks use the same formatting, but should be numbered separately, so a document may contain Theorem 1, Remark 1 and Example 1.

The text continues here. Proofs must be formatted as follows:

**Proof of Theorem 1.** Text of the proof. Note that the phrase “of Theorem 1” is optional if it is clear which theorem is being referred to. Always finish a proof with the following symbol. □

The text continues here.

4. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

5. Conclusions

This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.

6. Patents

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

**Institutional Review Board Statement:** In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving humans. OR “The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving animals. OR “Ethical review and approval were waived for this study due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals.

**Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable.” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

**Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics. If the study did not report any data, you might add “Not applicable” here.

**Acknowledgments:** In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results”.

**Appendix A**

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Appendix B**

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled starting with “A”—e.g., Figure A1, Figure A2, etc.

References

References must be numbered in order of appearance in the text (including citations in tables and legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as EndNote, ReferenceManager or Zotero to avoid typing mistakes and duplicated references. Include the digital object identifier (DOI) for all references where available.

Citations and references in the Supplementary Materials are permitted provided that they also appear in the reference list here.

In the text, reference numbers should be placed in square brackets [ ] and placed before the punctuation; for example [1], [1–3] or [1,3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 10), or [6] (pp. 101–105).

1. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name* **Year**, *Volume*, page range.
2. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.
3. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008; pp. 154–196.
4. Author 1, A.B.; Author 2, C. Title of Unpublished Work. *Abbreviated Journal Name* year, *phrase indicating stage of publication (submitted; accepted; in press)*.
5. Author 1, A.B. (University, City, State, Country); Author 2, C. (Institute, City, State, Country). Personal communication, 2012.
6. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference (Day Month Year).
7. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.
8. Title of Site. Available online: URL (accessed on Day Month Year).