**Relazione Progetto**

**Internet of Things based smart systems**

**Valutazione dell'Impatto della Riduzione della Precisione dei Pesi in una Rete Neurale nei Confronti dell'Accuratezza di Classificazione**

**Alessandro Messina, matricola O55000354**

**Orazio Scavo, matricola O55000414**

ANNO ACCADEMICO 2018/2019

Sommario

***[1](#_Toc535363772)******[Introduzione](#_Toc535363772)*** [- 3 -](#_Toc535363772)

[**1.1** **Ambito applicativo** - 3 -](#_Toc535363773)

[**1.2** **Obiettivi** - 3 -](#_Toc535363774)

[**1.3** **Flusso dell’analisi** - 4 -](#_Toc535363775)

[***2*** ***Rete originale*** - 5 -](#_Toc535363776)

[**2.1** **Progettazione della rete neurale** - 5 -](#_Toc535363777)

[**2.2** **Scelte operative** - 5 -](#_Toc535363778)

[**2.3** **Valutazione dell’accuratezza della rete** - 5 -](#_Toc535363779)

[***3*** ***Approximate Computing sulla rete neurale realizzata*** - 6 -](#_Toc535363780)

[**3.1** **Standard IEEE sulla rappresentazione dei float** - 6 -](#_Toc535363781)

[**3.2** **Scelte delle configurazioni approssimate** - 6 -](#_Toc535363782)

[**3.2.1** **Layer approssimati** - 7 -](#_Toc535363783)

[**3.2.2** **Entità dell’approssimazione** - 7 -](#_Toc535363784)

[**3.3** **Flusso dell’approssimazione** - 7 -](#_Toc535363785)

[**3.4** **NNAXIM Tool (Neural Network Approximate Computing SIMulator)** - 8 -](#_Toc535363786)

[**3.4.1** **Reti disponibili** - 8 -](#_Toc535363787)

[**3.4.2** **Struttura del progetto** - 8 -](#_Toc535363788)

[**3.4.3** **L’interfaccia e le operazioni disponibili** - 9 -](#_Toc535363789)

[***4*** ***Valutazione dei risultati ottenuti*** - 10 -](#_Toc535363790)

[***5*** ***Conclusioni*** - 11 -](#_Toc535363791)

[**Bibliografia** - 12 -](#_Toc535363792)

# ***Introduzione***

Negli ultimi anni l’Approximate Computing è stato riscoperto come uno degli approcci più promettenti per la riduzione del consumo di energia in molte applicazioni caratterizzate da una certa tolleranza agli errori (si dice che queste applicazioni abbiano una *forgiving nature*). Questa tendenza è sicuramente legata alla crescente importanza che i consumi energetici assumono nella maggior parte dei dispositivi in uso al giorno d’oggi (dispositivi mobili, IoT, data center, etc.). La riduzione della quantità di energia e risorse richieste da parte di questi dispositivi comporta una serie di vantaggi non trascurabili, quale ad esempio nel caso di dispositivi IoT o mobili l’incremento della propria autonomia.

Uno degli ambiti che meglio si presta all’applicazione dell’Approximate Computing e quello delle *reti neurali,* in quanto spesso caratterizzate da una certa tolleranza agli errori, ma soprattutto capaci di auto-correggere gli errori introdotti dalle approssimazioni (*self error-healing*), grazie alla possibilità di eseguire ulteriori training, successivi all’approssimazione, dei parametri della rete. Con il retraining della rete, i pesi approssimati riescono a convergere verso dei valori tali da consentire una minimizzazione dell’effetto dell’errore introdotto dall’approssimazione e una conseguente massimizzazione dei vantaggi della stessa.

Uno dei principali problemi che nascono quando si lavora con delle reti neurali, avendo a disposizione delle risorse limitate, riguarda la quantità di memoria utilizzata per memorizzare i parametri della rete. Ridurre ad esempio il numero di bit per rappresentare tali pesi ha un impatto molto positivo sulla quantità di risorse richieste. La riduzione del numero di bit offre inoltre la possibilità di utilizzare circuiti aritmetici ridotti con una conseguente riduzione dell’area, potenza e possibilmente una riduzione del percorso critico e quindi un aumento della frequenza di clock. Una tale approssimazione ha però delle conseguenze anche sull'accuratezza della rete neurale. Lo scopo di questa trattazione è proprio quello di valutare tale impatto.

## **Ambito applicativo**

Per rendere più realistica l’analisi, è stata pensata una specifica applicazione in ambito IoT per la gestione di un parcheggio, in cui è richiesta la classificazione degli oggetti immortalati da una fotocamera posta all’entrata dello stesso, in modo da identificare diverse tipologie di veicolo o altri oggetti, la cui presenza in quel punto rappresenta una situazione non prevista. Per questo tipo di applicazione (classificazione) viene solitamente utilizzata una rete neurale di tipo convoluzionale (CNN), che data in input una certa immagine è capace di rilevare oggetti al suo interno e classificarli associandoli ad una tra un certo numero di categorie.

## **Obiettivi**

L’obiettivo principale di questa trattazione è quello di valutare il *tradeoff* tra perdita di accuratezza e memoria risparmiata in seguito ad una riduzione del numero di bit utilizzati per rappresentare i pesi dei neuroni della rete. A tal fine, sono state scelte 9 configurazioni, variabili sulla base di dove l’approssimazione è applicata e in che misura, e per ognuna di esse è stata effettuata la valutazione di cui sopra.

Per rendere più agevole il processo di training e test delle varie configurazioni è stata inoltre implementata una semplice applicazione in C++, NNAXIM [[1]](#footnote-1), che consente di analizzare il comportamento di una o tutte le configurazioni. Per la realizzazione del programma è stata utilizzata la libreria *TinyDNN[[2]](#footnote-2)*, che consente di implementare (sempre in C++) delle reti neurali a livello software.

## **Flusso dell’analisi**

I capitoli successivi approfondiscono i vari passi del flusso di lavoro seguito per l’analisi, il quale si articola nei seguenti punti:

* *Capitolo 2*. Progettazione, allenamento e valutazione dell’accuratezza della rete originale.
* *Capitolo 3*. Applicazione dell’Approximate Computing sulla rete realizzata.
* *Capitolo 4*. Confronto dei risultati ottenuti per le diverse configurazioni.
* *Capitolo 5*. Conclusioni sul lavoro svolto.

# ***Rete originale***

Data l’applicazione, è stato necessario trovare una rete neurale capace di effettuare la classificazione richiesta a partire dalle immagini di input prodotte dalla fotocamera del sistema. Una volta trovata la rete è stato necessario allenarla per ottenere i valori dei pesi per la configurazione base da usare successivamente come punto di partenza per l’applicazione dell’approssimazione nelle varie configurazioni. È stata inoltre valutata l’accuratezza della rete originale in modo da poter effettuare il confronto con le configurazioni approssimate.

## **Progettazione della rete neurale**

Per la classificazione è stata scelta una *Convolutional Neural Network* utilizzata all’interno dell’esempio *cifar10* di TinyDNN. Questa rete è composta da una serie di 3 blocchi di layer (ognuno dei quali costituito da un layer convoluzionale, un pooling layer e un relu activation layer) seguiti da un fully-connected layer, un relu activation layer e infine un ulteriore fully-connected layer. Per quanto riguarda i dataset di training e di test sono stati usati quelli di CIFAR10[[3]](#footnote-3). Questi contengono delle immagini 32x32px raffiguranti degli oggetti riconducibili ad una di queste 10 categorie: aeroplani, automobili, uccelli, gatti, cervi, cani, rane, cavalli, navi e camion. La rete neurale grazie al training impara quindi classificare gli oggetti presenti nelle immagini di input alla rete in una delle suddette categorie.

Tramite l’interfaccia offerta dalla libreria TinyDNN è stato possibile implementare le operazioni di training, test, caricamento dei dataset e quant’altro.

## **Scelte operative**

La rete progettata è caratterizzata da un’elevata complessità. Tale complessità si riflette sui tempi di allenamento della stessa e non avrebbe quindi consentito di lavorare agevolmente su di essa durante le fasi dello sviluppo del tool di simulazione. Al fine di disaccoppiare le fasi di sviluppo del tool e di training della rete neurale non approssimata, si è scelto inizialmente di lavorare su una rete estremamente più semplice, il cui obiettivo è quello di predire l’output di una funzione sinusoidale[[4]](#footnote-4). In questo modo è stato possibile sviluppare e testare velocemente le procedure di training e test sulla rete e la definizione delle varie configurazioni da applicare per l’approssimazione. Dopo aver verificato il corretto funzionamento delle procedure sviluppate è stato effettuato un *porting* dell’applicazione sul modello di rete descritto nel paragrafo precedente. L’applicazione è stata poi sfruttata per eseguire una sola volta il training e il test delle reti approssimate (e non) su un calcolatore più performante, riducendo così le latenze di esecuzione.

## **Valutazione dell’accuratezza della rete**

L’accuratezza di una rete neurale usata per la classificazione può essere facilmente valutata basandosi sulla percentuale di classificazioni corrette su quelle totali effettuate sul dataset di test.

# ***Approximate Computing sulla rete neurale realizzata***

Come già detto nei capitoli precedenti, l’approximate computing applicato alla rete riguarda l’approssimazione dei bit usati per memorizzare i pesi della rete. Nell’approccio ideale sarebbe necessario applicare approssimazioni successive in maniera iterativa (approssimazione, test, training, test) fino a trovare il giusto tradeoff tra quantità di memoria risparmiata e accuratezza persa. Fintanto che l’accuratezza rimane sopra il requisito minimo fornito dall’applicazione è possibile continuare ad approssimare. A causa della carenza di risorse computazionali a disposizione si è scelto di utilizzare solamente 9 configurazioni di approssimazione prestabilite.

## **Standard IEEE sulla rappresentazione dei float**

TODO - Orazio.

|  |  |  |
| --- | --- | --- |
| **x** | **x** | **x** |
| 9 | 18 | 12 |

Tabella 1. Errore introdotto approssimando i bit di rappresentazione dei float secondo lo standard IEEE 754

## **Scelte delle configurazioni approssimate**

Nella Tabella 2 è mostrata la lista delle configurazioni delle approssimazioni adottate. Si può notare come ognuna delle 9 configurazioni è identificata dai layer coinvolti nell’approssimazione e dalla misura con la quale questa è applicata ai pesi dei neuroni di questi layer.

|  |  |  |
| --- | --- | --- |
| **Configurazione** | **Numero di bit di ogni peso dei neuroni dei layer di input e output** | **Numero di bit di ogni peso dei neuroni degli hidden layer** |
| Originale | 32 | 32 |
| 1 | 22 | 22 |
| 2 | 18 | 18 |
| 3 | 14 | 14 |
| 4 | 32 | 22 |
| 5 | 32 | 18 |
| 6 | 32 | 14 |
| 7 | 22 | 18 |
| 8 | 22 | 14 |
| 9 | 18 | 12 |

Tabella 2. Configurazioni scelte per le approssimazioni del numero di bit dei pesi.

A seguire sono spiegate le motivazioni della scelta di queste configurazioni.

### **Layer approssimati**

La prima cosa sulla quale è importante soffermarsi riguarda la tipologia di layer sui quali è stata applicata l’approssimazione. Le 9 configurazioni viste in Tabella 2 possono essere suddivise a tal proposito in 3 gruppi:

* *Gruppo 1*; approssimazione omogenea sui neuroni di tutti i layer della rete.
* *Gruppo 2*; approssimazione applicata solo agli hidden layer della rete.
* *Gruppo 3*; approssimazione differente tra neuroni degli hidden layer e neuroni dei layer di input e output.

Questa classificazione è stata effettuata basandosi sui risultati riportati in [1], [2] e [3], i quali, tramite test più esaustivi, hanno dimostrato che i neuroni degli hidden layer possiedono una maggiore resilienza agli errori, viceversa i neuroni dei layer di input e output sono più sensibili.

### **Entità dell’approssimazione**

Ad ognuno dei gruppi di cui sopra appartengono 3 configurazioni che differiscono tra loro in base al numero di bit utilizzati per l’approssimazione. Tale quantità è stata scelta sulla base della Tabella 1. Dalla tale tabella risulta evidente come approssimazioni inferiori o uguali ai 10 bit (e che quindi codificano i float con un numero di bit maggiore o uguale a 22) introducono un errore praticamente trascurabile se le grandezze in gioco sono dell’ordine dell’unità o più, mentre approssimazioni maggiori o uguali di 20 bit (e che quindi codificano i float con un numero di bit minore o uguale a 12) introducono un errore molto grande (a prescindere dell’ordine delle grandezze in gioco). Per questo motivo, in tutte le configurazioni, l’approssimazione minima introdotta è di 10 bit, mentre la massima è di 20 bit. Ridurre la dimensione di ogni peso della rete di almeno 10 bit comporta un risparmio in termini di memoria molto elevato a fronte di una perdita di accuratezza molto bassa.

## **Flusso dell’approssimazione**

Dopo aver mostrato le configurazioni scelte per le approssimazioni e le motivazioni dietro di esse è necessario spiegare come ogni approssimazione viene applicata alla rete. Il flusso dell’approssimazione è uguale per ognuna delle configurazioni e prevede oltre all’approssimazione in sé, anche delle fasi di test e retraining della rete. In particolare, vengono effettuate le seguenti fasi:

1. *Approssimazione dei pesi secondo la configurazione scelta*. Ogni peso della rete viene approssimato (e non troncato, al fine di minimizzare l’errore introdotto) al numero di bit stabilito dalla configurazione. In Fig.1 è mostrato il pezzo di codice (C++) utilizzato all’interno di NNAXIM per applicare tale approssimazione. La configurazione è definita tramite i due parametri hidden\_nlayer\_bitseextern\_nlayer\_bitsche indicano rispettivamente il numero di bit da utilizzare per i neuroni degli hidden layer e il numero di bit da utilizzare per i neuroni dei layer di input e output. La funzione che applica effettivamente l’approssimazione sul singolo peso è roundb(…). Durante questa fase viene raccolto anche il numero di bit risparmiati grazie all’approssimazione.
2. *Test dopo l’approssimazione*. La rete viene testata in modo da individuare la perdita di accuratezza rispetto alla configurazione originale.
3. *Retraining*. Per mitigare l’effetto negativo sull’accuratezza dovuto all’approssimazione viene sfruttata la capacità di self-healing della rete neurale tramite un opportuno training. Questo training non necessità dello stesso numero di epoche del training della rete originale, poiché i pesi sono già allenati e sono quindi più veloci a convergere. Questa fase è fondamentale e molto delicata poiché il numero effettivo di epoche da eseguire è di cruciale importanza: è stato notato come un numero troppo basso ha effetti addirittura negativi sull’accuratezza della rete (anche di diversi ordini di grandezza), mentre un numero troppo alto è superfluo e comporta solo una perdita di tempo.
4. *Test dopo il retraining*. Questa fase ha un duplice scopo: capire se il numero di epoche del retraining è stato sufficiente e calcolare la perdità di accuratezza effettiva dell’aprossimazione rispetto alla configurazione originale.

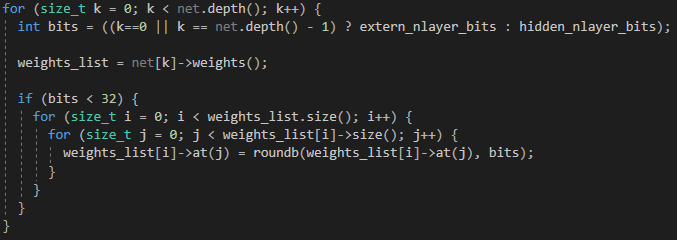


Figura 1. Codice per l'approssimazione dei bit dei pesi per una specifica configurazione.

## **NNAXIM Tool (Neural Network Approximate Computing SIMulator)**

NNAXIM (Neural Network Approximate Computing SIMulator) è un tool che consente di simulare l’applicazione dell’approximate computing sul numero di bit usati per rappresentare i pesi di una neural network. Il tool, disponibile su GitHub all’indirizzo <https://github.com/Taletex/NNAXIM>, è scritto in C++ e utilizza la libreria TinyDNN per l’implementazione delle reti neurali. Il tool è stato realizzato per facilitare le operazioni di training, test e raccolta delle prestazioni durante l’applicazione dell’AC sulle neural network.

### **Reti disponibili**

Nel tool è possibile utilizzare due reti: una *CNN* per la classificazione (rete di default) e una *MLP* per l’approssimazione di funzioni. Per cambiare rete è necessario cambiare i file sorgente che si trovano all’interno della directory *src/ac\_nn* con quelli presenti in *src/sinus\_fit\_version*.

### **Struttura del progetto**

Le directory del tool sono organizzate così come segue:

* doc; contiene la documentazione del progetto (incluso questo file)
* src; contiene i file sorgente del progetto. Al suo interno troviamo a sua volta diverse cartelle:
  + *ac\_nn*; contiene i file sorgente di NNAXIM che consistono in un main (*ac\_nn.cpp*), le librerie sviluppate per il tool (*ac\_nn\_lib.cpp* e *ac\_nn\_lib.hh*), i file con il modello e i pesi della rete (dentro *net\_params*) e un file di log (*log.txt*, usato per memorizzare i risultati dei test automatici). Dentro *ac\_nn* è utilizzata di default una CNN per la classificazione, motivo per il quale è possibile trovarvi anche la cartella *cifar-10-batches-bin* all’interno della quale sono presenti i dataset di test e training da usare con la CNN.
  + *sinus\_fit\_version*; contiene gli stessi file presenti in ac\_nn riadattati per una MLP per l’approssimazione di funzioni. È sufficiente sostituire i file presenti in ac\_nn con quelli presenti in sinus\_fit\_version per usare la MLP.
  + *build*; cartella che contiene (una volta buildato il progetto tramite cmake) la soluzione Visual Studio di NNAXIM.
  + *cereal, cmake, third\_party, tiny\_dnn*; folder provenienti dalla libreria TinyDNN e necessari per il suo utilizzo.

### **L’interfaccia e le operazioni disponibili**

Al primo avvio dell’applicazione, se non sono stati scaricati anche i pesi e il modello di default, sarà necessario eseguire un primo training della (Fig. 2) rete che richiederà parecchio tempo, a seconda del calcolatore utilizzato[[5]](#footnote-5).

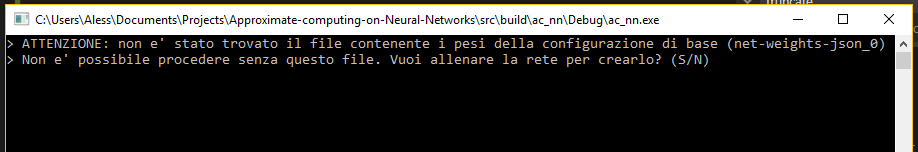


Figura 2. Schermata di avvio nel caso in cui non è presente il file contenente i pesi della rete originale.

Se i pesi della rete originale sono disponibili, all’avvio comparirà l’interfaccia base dell’applicazione (Fig. 3). Da questa è possibile effettuare principalmente 3 azioni:

1. Scegliere una configurazione sulla quale lavorare. In uno step successivo sarà possibile:
   1. indicare se si vuole effettuare un allenamento della rete con i pesi approssimati e successivo salvataggio degli stessi
   2. indicare se si vuole solo testare la rete con la configurazione scelta.
2. Eseguire il test automatico di tutte le configurazioni. Questa funzionalità, dopo un primo test della rete nella sua configurazione originale, farà partire 9 iterazioni (una per ogni configurazione) nelle quali saranno svolte le 4 fasi descritte in 3.3. Alla fine delle iterazioni, i risultati dei test saranno mostrati a video e saranno anche salvati su file (log.txt).
3. Uscire dall’applicazione.

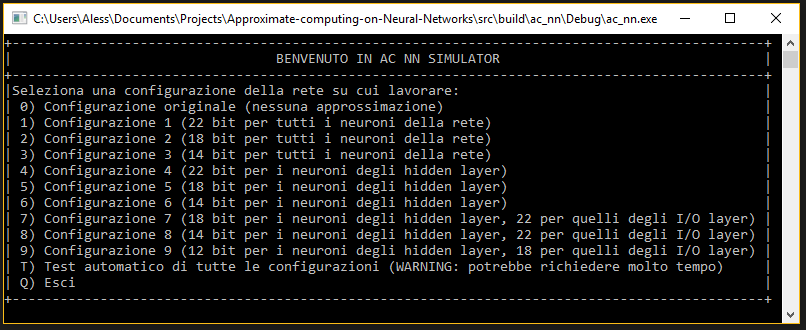


Figura 3. Schermata di avvio nel caso in cui è presente il file contenente i pesi della rete originale.

# ***Valutazione dei risultati ottenuti***

TODO

Per ognuna di queste configurazioni è stata valutata l’accuratezza prima e dopo dell’esecuzione di un riallentamento della rete (che consente in generale di mitigare l’impatto sulle prestazioni dovuto all’approssimazione dei pesi).

Una volta ottenuti tutti i valori di accuratezza per le varie configurazioni, questi sono stati graficati e confrontati opportunamente.

GRAFICI E RELATIVO COMMENTO

# ***Conclusioni***

TODO Conclusioni

# **Bibliografia**

|  |  |
| --- | --- |
| [1] | Zidong Du, Avinash Lingamneni, Yunji Chen, Krishna Palem, Olivier Temam, Chengyong Wu, "*Leveraging the Error Resilience of Machine-Learning Applications for Designing Highly Energy Efficient Accelerators*", 2014. |
| [2] | Swagath Venkataramani, Ashish Ranjan, Kaushik Roy and Anand Raghunathan, "*AxNN: Energy-Efficient Neuromorphic Systems using*  *Approximate Computing*", 2014. |
| [3] | Qian Zhang, Ting Wang, Ye Tian, Feng Yuan and Qiang Xu, "*ApproxANN: An Approximate Computing Framework for Artificial Neural Network",* 2015. |

1. NNAXIM (Neural Network Approximate Computing SIMulator) è disponibile su GitHub presso l’indirizzo <https://github.com/Taletex/NNAXIM>. Il nome del progetto è un tributo al simulatore NOXIM, disponibile sempre su GitHub all’indirizzo <https://github.com/davidepatti/noxim>. [↑](#footnote-ref-1)
2. TinyDNN è disponibile gratuitamente su GitHub presso l’indirizzo <https://github.com/tiny-dnn/tiny-dnn>. [↑](#footnote-ref-2)
3. <https://www.cs.toronto.edu/~kriz/cifar.html>. [↑](#footnote-ref-3)
4. Come base di partenza per l’implementazione di questa rete è stato utilizzato l’esempio *sinus\_fit* offerto da TinyDNN. [↑](#footnote-ref-4)
5. Il tool è stato testato solo su Windows ed è necessario utilizzare Visual Studio 2015 o versioni successive e Cmake per poterlo compilare. [↑](#footnote-ref-5)