**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. Introduzione*** - 3 -](#_Toc535319182)

[**1.1 Ambito applicativo** - 3 -](#_Toc535319183)

[**1.3 Flusso dell’analisi** - 3 -](#_Toc535319184)

[***2. Rete originale*** - 5 -](#_Toc535319185)

[**2.1 Progettazione della rete neurale** - 5 -](#_Toc535319186)

[**2.2 Allenamento della rete neurale** - 5 -](#_Toc535319187)

[**2.3 Valutazione dell’accuratezza della rete** - 5 -](#_Toc535319188)

[***3. Approximate Computing sulla rete neurale realizzata*** - 6 -](#_Toc535319189)

[**3.1 Scelta delle configurazioni approssimate** - 6 -](#_Toc535319190)

[***4. Valutazione dei risultati ottenuti*** - 7 -](#_Toc535319191)

[***5. Conclusioni*** - 8 -](#_Toc535319192)

# ***1. Introduzione***

🡪 In questa descrizione spiegare perché stiamo usando AC nelle NN

La quantità di memoria utilizzata per memorizzare i parametri di una rete neurale è dominata dalla memorizzazione dei pesi. Ridurre il numero di bit per rappresentare i pesi ha quindi un impatto positivo sull'utilizzo 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. La riduzione del numero di bit utilizzati per rappresentare i pesi ha sicuramente un impatto sull'accuratezza della rete neurale. Si vuole valutare tale impatto.

Modificare questa breve introduzione che ha scritto il PROF e parlare di:

* Error tolerance e Forgiving nature delle NN.
* Self error healing delle NN.

## **1.1 Ambito applicativo**

* Descrizione ambito applicativo (parcheggio, fotocamera, etc.).
  + Descrizione dell’applicazione per la quale è stata pensata la rete che stiamo cercando di ottimizzare (in termini di memoria). RICORDA: IOT.
* Spiegare in 2 parole precise perché usiamo una CNN (perché famo classificazione). Nota: la spiegazione completa e tecnica sarà fornita in 2.1. Qui spiega solo l’applicazione!

**1.2 Obiettivi**

Spiegazione obiettivi del progetto.

* Parametri
* Citare il programmino per scegliere la configurazione della rete da testare.
  + Realizzare il programmino in R per graficare i risultati finali.

## **1.3 Flusso dell’analisi**

I capitoli successivi spiegano i vari passi del flusso di lavoro adottato, il quale si articola nei seguenti punti:

1. Progettazione, allenamento e valutazione accuratezza della rete originale
   1. Progettazione della rete neurale.
      1. Spiegazione della rete che dovrà usare il dataset CIFER-10.
   2. Allenamento della rete neurale nella sua configurazione originale
   3. Valutazione dell'accuratezza della rete (Aorig)
2. Approximate Computing sulla rete realizzata
   1. Iterazione orizzontale (ovvero 2/3 iterazione max in cui applichiamo l’approssimazione a insiemi di neuroni diversi – e.g. tutti i neuroni, solo i neuroni hidden solo alcuni dei neuroni hidden)
      1. Iterazione verticale (ovvero 2/3 iterazioni max in cui applichiamo in maniera diversa l’approssimazione ai neuroni dell’insieme dell’iterazione di livello superiore, ovvero con un numero di bit di rappresentazione dei pesi diverso)
         1. Modifica dei pesi riducendo i bit di rappresentazione
         2. Valutazione dell'accuratezza della rete (Amod)
         3. Riallenamento la rete in cui i pesi sono rappresentati con un numero ridotto di bit (2 epoche max)
         4. Valutare l'accuratezza della rete (Amod2)

* Ps. Fare riferimento ai paper quando scegli le configurazioni con solo gli hidden layer
* Ps2. Fare notare le problematiche di progetto: a causa delle tempistiche del training abbiamo deciso di scegliere 3x3 iterazioni (orizzontali e verticali rispettivamente) e non un numero “quanto basta” come descritto nei paper.
  + Es. In letteratura sono state descritte diverse soluzioni per applicare l’AC sulle NN. Noi abbiamo seguito un approccio ibrido cercando di approssimare per step successivi sia a livello orizzontale che a livello verticale. A causa delle limitate risorse hardware a disposizione non abbiamo però potuto eseguire un numero esaustivo di iterazioni, motivo per il quale abbiamo deciso di adottare un numero di iterazioni fisso nelle quali vengono utilizzate le configurazioni che sono state considerate più significative per la valutazione dei risultati finali.

1. Confrontare Aorig, Amod e Amod2 (confronto generale di tutte le configurazioni adottate)
2. Conclusioni

* Bibliografia

# ***2. Rete originale***

## **2.1 Progettazione della rete neurale**

Xxx

## **2.2 Allenamento della rete neurale**

Allenamento della rete neurale

* Spiegazionde uso cloud

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

Asd

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

## **3.1 Scelta delle configurazioni approssimate**

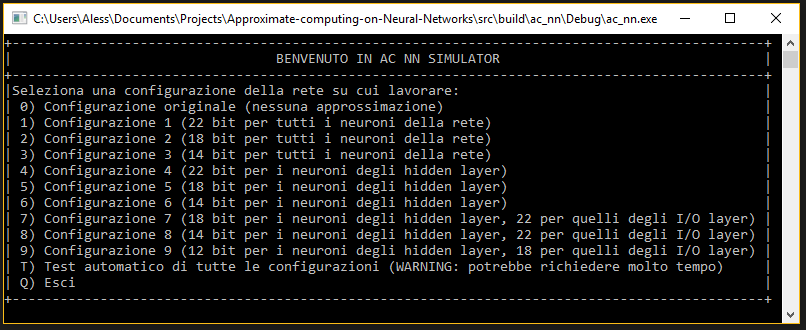
Xxx

## **3.2 AXIM Tool (AC simulator for NN)**

- Abbiamo reliazzato il programma per poter studiare la rete

- Descrizione directories del progetto (tipo dove trovi i file sorgenti, etc.)

- Come funziona il tool e cosa permette di fare



# ***4. Valutazione dei risultati ottenuti***

# ***5. Conclusioni***