Voice Commands for Robot Orientation in Space: A TinyML Approach using Transfer Learning

Oneata Razvan-Mihai, Troaca Viorel Mario and **Ghimpeteanu Alexandru Andrei**
Universitatea din București
razvan-mihai.oneata@s.unibuc.ro
viorel-mario.troaca@s.unibuc.ro
alexandru-andrei.ghimpeteanu@s.unibuc.ro

Abstract

Acest document prezintă dezvoltarea și validarea unui sistem de control vocal pentru roboți mobili, utilizând tehnici avansate de TinyML. Folosind platforma Edge Impulse, am implementat un model bazat pe *Transfer Learning* (MobileNetV2) capabil să clasifice comenzi de orientare ("Left", "Right", "Go", "Stop") cu o acuratețe de 93.2%. Noutatea abordării constă în optimizarea modelului pentru resurse limitate și validarea robusteții acestuia la zgomot ambiental prin utilizarea unui set de date hibrid. Rezultatele demonstrează o latență de inferență redusă (aprox. 59 ms), permițând integrarea viitoare cu un sistem de navigație inerțială (IMU) pentru control în buclă închisă.

1 Introducere

Interacțiunea om-robot (HRI) naturală este esențială pentru adoptarea roboților în mediile casnice și industriale. Metodele tradiționale, bazate pe telecomenzi, limitează operabilitatea. Propunem un sistem "Voice Commands for Robot Orientation in Space", care permite controlul direcției robotului prin comenzi vocale procesate local (*on-device*).

Obiectivul principal este eliminarea latenței specifice procesării în cloud și asigurarea confidențialității datelor. Soluția noastră utilizează paradigma TinyML, rulând inferența direct pe dispozitive embedded. Deși integrarea finală vizează un controler PID și senzori IMU pentru precizie mecanică, acest raport se concentrează pe arhitectura, antrenarea și validarea performanței modulului de recunoaștere vocală (*Keyword Spotting*).

2 Lucrări Conexe

Conceptul de *TinyML* a revoluționat robotica mobilă, permițând rularea rețelelor neurale pe microcontrollere (MCU) cu consum redus de energie. Warden (2018) a stabilit standardul în domeniu prin lansarea setului de date *Google Speech Commands*, facilitând antrenarea modelelor supravegheate.

În timp ce Sainath and Parada (2015) au demonstrat eficiența Rețelelor Neurale Convoluționale (CNN) pentru sarcini audio, antrenarea acestora de la zero pe seturi de date limitate poate duce la *overfitting* sau performanțe sub-optime. Studii recente arată că tehnicile de *Transfer Learning*, aplicate pe arhitecturi precum MobileNet, oferă un echilibru superior între acuratețe și cost computațional pentru aplicații embedded.

3 Metodologie

Metodologia abordată urmărește întregul flux de lucru TinyML: achiziția datelor, procesarea semnalului (DSP), antrenarea modelului și evaluarea performanței.

3.1 Pregătirea Setului de Date

Am utilizat o versiune optimizată a setului de date *Google Speech Commands* (versiunea "Mini"), concentrându-ne pe patru clase de control: "Left", "Right", "Go", "Stop".

Pentru a asigura robustețea sistemului în scenarii reale, am construit un set de date hibrid:

- **Clase Principale:** Eșantioane de 1 secundă din setul Google.
- **Clasa Unknown:** Compusă din cuvinte auxiliare ("Up", "Down", "Yes", "No") pentru a reduce declanșările false.
- **Clasa Noise (Custom):** Am înregistrat manual 5 minute de zgomot ambiental (liniște, sunet de fundal) folosind hardware-ul de test, pentru a preveni activarea robotului în absența comenzilor.

În total, setul de date a fost echilibrat folosind tehnica de *auto-weighting* în timpul antrenamentului, pentru a compensa disproporția dintre clasa "Unknown" și comenzile specifice.

3.2 Procesarea Semnalului (MFE)

Semnalul audio brut este procesat folosind blocuri MFE (*Mel-filterbank Energy*), care sunt mai eficiente computațional decât MFCC pe hardware limitat.

- **Fereastra de timp:** 1000 ms.
- **Stride:** 500 ms (suprapunere pentru detectarea continuă).
- **Frecvență:** 16000 Hz.

Acest proces transformă sunetul în spectrograme care servesc drept imagini de intrare pentru rețeaua neurală.

3.3 Arhitectura Modelului: Transfer Learning

Inițial, am experimentat cu o arhitectură CNN simplă (2 straturi convoluționale), însă aceasta a prezentat o capacitate redusă de învățare (acurațete sub 50%) din cauza complexității clasei "Unknown".

Soluția finală adoptată utilizează Transfer Learning bazat pe modelul MobileNetV2 0.35. Această rețea este pre-antrenată pe seturi de date masive (ImageNet/AudioSet) și re-antrenată (fine-tuned) pe datele noastre specifice. Parametrii de antrenare:

- **Epoci:** 30.
- **Learning Rate:** 0.05.
- **Data Augmentation:** Activat (adăugare zgomot aleatoriu).

4 Rezultate Experimentale

Evaluarea modelului s-a realizat pe un set de testare separat (20% din date), iar performanța în timp real a fost validată prin emulare WebAssembly pe dispozitive mobile.

4.1 Metrice de Performanță ale Modelului

Modelul MobileNetV2 a obținut o acurațete globală de 93.2% pe setul de validare, cu o pierdere (loss) de 0.24. Matricea de confuzie relevă performanțe excelente pe clasele critice de siguranță:

- **Right:** 94.2% precizie.
- **Left:** 92.3% precizie.
- **Stop:** 91.1% precizie.
- **Noise:** 96.6% precizie (demonstrând rezistența la declanșări false în liniște).

Singura confuzie minoră observată este între clasa "Go" (82.9%) și "Unknown", o eroare acceptabilă care nu pune în pericol siguranța robotului (False Negative).

4.2 Eficiență Computațională

Pentru a estima fezabilitatea rulării pe un microcontroller (ex. Arduino Nano 33 BLE sau ESP32), am analizat resursele necesare modelului cuantizat (int8):

- **Timp de inferență estimat:** 59 ms (mult sub limita de 500 ms propusă inițial).
- **Memorie RAM (Peak):** 168.7 KB.
- **Flash Usage:** 693.2 KB.

4.3 Validare în Timp Real

Deoarece implementarea fizică pe șasiul robotului este în curs de desfășurare, validarea funcțională s-a realizat prin portarea modelului pe un smartphone via browser (WebAssembly). Testele live au confirmat latența insesizabilă și capacitatea modelului de a distinge clar între comenzile vocale și zgomotul de fundal într-un mediu de laborator.

5 Concluzie

Am dezvoltat cu succes un sistem de recunoaștere vocală robust pentru orientarea roboților, utilizând Edge Impulse și Transfer Learning. Trecerea la arhitectura MobileNetV2 a permis atingerea unei acurațete de peste 93%, superioară abordărilor clasice CNN simpliste.

References

Edge impulse. https://www.edgeimpulse.com.

Tara N Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.