Progetto Machine Learning e Sistemi Intelligenti per Internet

Studente: Massimo D'Alessandro

Matricola: 563758

GitHub progetto: https://github.com/massidale/instrument-classifier.git

Questo progetto consiste nell'addestramento di una Rete Neurale Convoluzionale (CNN) che si prefissa come obiettivo quello di riconoscere lo strumento musicale principale all'interno di una traccia audio.

Quando si ascolta uno strumento musicale, ciò che consente di riconoscerlo all'orecchio sono diversi fattori, e alcuni di questi possono essere tradotti in modo che siano comprensibili da una macchina. In particolare si hanno il timbro, l'inviluppo e la gamma di frequenze raggiunte dallo strumento.

Timbro
Quando viene suonata una nota, è possibile ottenere molte informazioni di questa nel dominio delle frequenze. La caratteristica principale che permette di distinguere uno strumento dall'altro è il timbro: studiando la trasformata di Fourier di una specifica nota si nota come la frequenza fondamentale si mantiene la stessa per ogni strumento usato, mentre ciò che cambia sono le armoniche, ovvero le frequenze che si trovano in multipli interi della frequenza portante; queste sono quelle che stabiliscono il timbro dello strumento.

Inviluppo
Un'altra discriminante è l'inviluppo ASDR dello strumento, osservato nel dominio del tempo, che è suddiviso in 4 parti:

1. Attack: la fase in cui viene raggiunta l'ampiezza massima.
2. Decay: dopo l'attacco, l'ampiezza si stabilizza su un valore costante.
3. Sustain: fase durante la quale la nota è mantenuta costante.
4. Release: fase in cui la nota decresce fino ad estinguersi completamente.

Questo è diverso in ogni strumento.

Gamma di frequenze
La gamma di frequenze raggiunte da uno strumento musicale è un elemento determinante per il suo riconoscimento, poiché ogni strumento ha un intervallo di frequenze specifico che può produrre. Alcuni strumenti hanno una gamma limitata (come il flauto dolce), altri una gamma intermedia (come violino o pianoforte), e altri ancora una gamma molto estesa. Questa caratteristica influenza non solo le note che lo strumento può suonare, ma anche la sua funzione in un ensemble o orchestra.

Strumenti utili per l'addestramento della rete
Per sfruttare al meglio le caratteristiche di una traccia musicale è necessario avere gli strumenti adatti per risaltarne gli aspetti critici: per osservare l'inviluppo è stato utilizzato il waveplot, che rappresenta l'intensità (o ampiezza) del segnale nel tempo. Il discorso è più ampio quando si parla di frequenze e timbro, poiché queste sono caratteristiche che vengono studiate nel dominio delle frequenze ed è necessario individuare la trasformata adatta allo scopo. Tuttavia una semplice trasformata non conserva l'informazione temporale, per questo si è deciso di utilizzare lo spettrogramma: uno strumento che consente di effettuare una qualsiasi trasformata in un intervallo di tempo ristretto, in modo da avere le informazioni in frequenza conservando le informazioni temporali.

Esistono diversi tipi di trasformate, in particolare per questo progetto sono state usate quella di Mel, CQT e cromogramma:

La scala di Mel è una scala ottimizzata per dare maggior importanza alle componenti udibili all'orecchio umano, quindi "amplificando" le frequenze centrali nello spettro udibile e comprimendo gradualmente quelle negli estremi, fino a tagliare quelle fuori.

La CQT utilizza una scala logaritmica, il che la rende adatta a rappresentare contenuti musicali, in particolare con diversi strumenti, dove intervalli come le ottave seguono una progressione logaritmica (un'ottava è l'intervallo tra due suoni in cui la frequenza di uno è esattamente il doppio di quella dell'altro). Ad esempio, tra 1000 e 2000 Hz c'è lo stesso "peso" di dettaglio che tra 100 e 200 Hz.

Il cromogramma invece è molto utile per identificare il timbro di un singolo strumento: questo utilizza la scala logaritmica ma invece di mappare le frequenze distintamente le mappa su 12 bin, ognuno corrispondente ad una singola nota (che sono 12, considerando toni e semitoni). Questo consente di avere un maggiore focus sulla frequenza principale e sulle armoniche, consentendo di identificare al meglio lo strumento utilizzato.

Dataset
Il dataset IRMAS (https://www.upf.edu/web/mtg/irmas) si compone di 6705 tracce musicali di 3 secondi, campionate a 44.1kHz estratte da canzoni, dove per ogni canzone si ha un massimo di 3 tracce. Il dataset è composto da 11 strumenti diversi (e quindi 11 classi) ma allo scopo di questo progetto sono state prese 3 classi: chitarra elettrica, pianoforte e sassofono. Queste hanno un elevato di campioni (760 per la chitarra elettrica, 721 per il piano e 680 per il sassofono) e sono strumenti con un timbro ben riconoscibile.

Preprocessamento dei dati
Si è controllato che tutte le tracce avessero durata di esattamente 3 secondi e che la frequenza di campionamento fosse di 44100hz per ogni traccia, dopodiché si è esplorato il dataset, ascoltando alcune tracce e osservando waveplot e vari tipi di spettrogrammi.

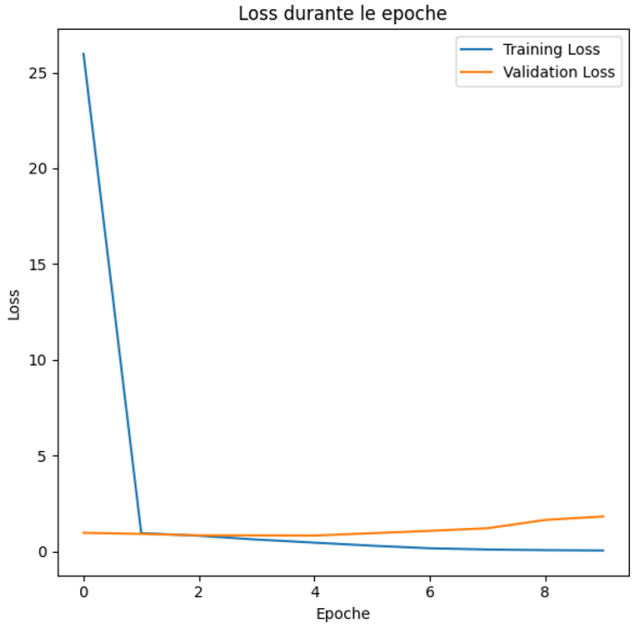
Tra i tentativi effettuati c'è stato anche quello di filtrare tutte le tracce high-pass (eliminando così i disturbi di sottofondo), e utilizzare la FFT come input per la rete neurale. Tuttavia la scelta finale è ricaduta su spettrogrammi che utilizzano delle trasformate specifiche (visti in precedenza) che danno maggiore importanza alle basse frequenze.

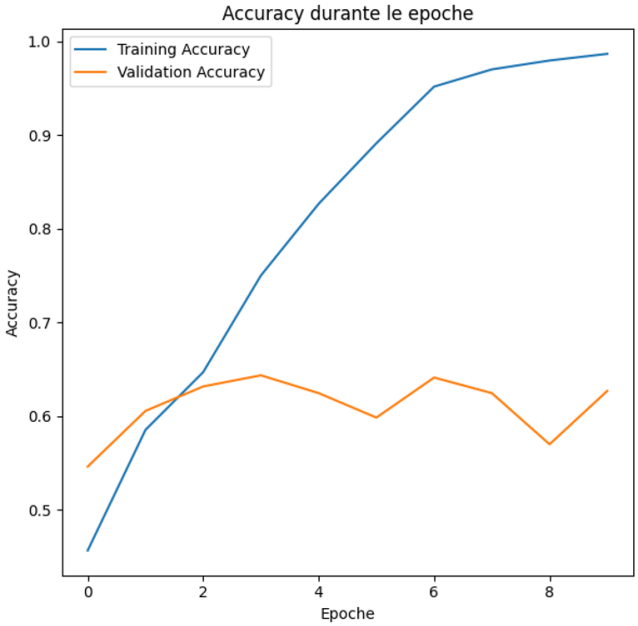
Struttura della rete
La rete neurale è stata costruita per accettare due input, un waveplot e uno spettrogramma (sono poi stati addestrati 3 modelli utilizzando i 3 diversi spettrogrammi); entrambi gli input vengono processati da due strati convoluzionali composti rispettivamente da 32 e 64 filtri 3x3, intervallati da due strati di max pooling con una maschera 2x2. Questi dati vengono poi ridotti ad una sola dimensione, concatenati e usati come input di uno strato denso con 128 neuroni e il 30% di dropout, seguito da uno strato di output per la classificazione dello strumento. La funzione di attivazione utilizzata è stata la ReLU per tutti gli strati tranne quello di output, nel quale è stata utilizzata la softmax per produrre una probabilità normalizzata per ogni classe.

Le etichette di ogni classe sono state codificate con la one hot encoding per evitare che la rete neurale interpreti i valori numerici delle etichette come aventi un significato intrinseco o un ordine gerarchico, il che potrebbe portare a un apprendimento distorto.

L'addestramento è stato effettuato utilizzando una k-fold cross validation con 5 fold, ciascuno addestrato per 10 epoche. Questo perché l'accuratezza sul training set raggiunge valori prossimi al

100% già intorno alla 6ª o 7ª epoca, mentre l'accuratezza sul validation set si stabilizza senza miglioramenti significativi, sintomo di un forte overfitting dovuto alla scarsità di dati.

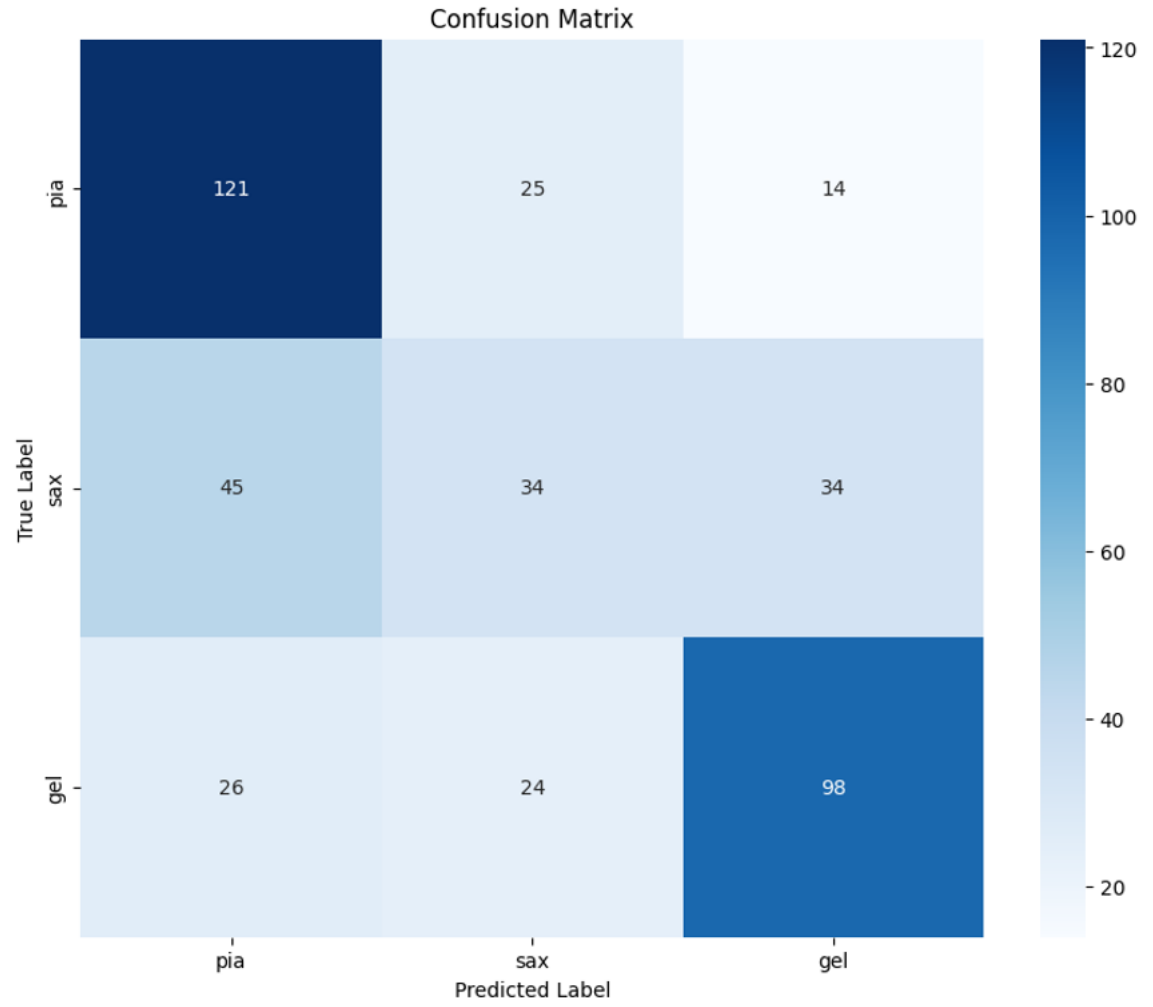




Risultati

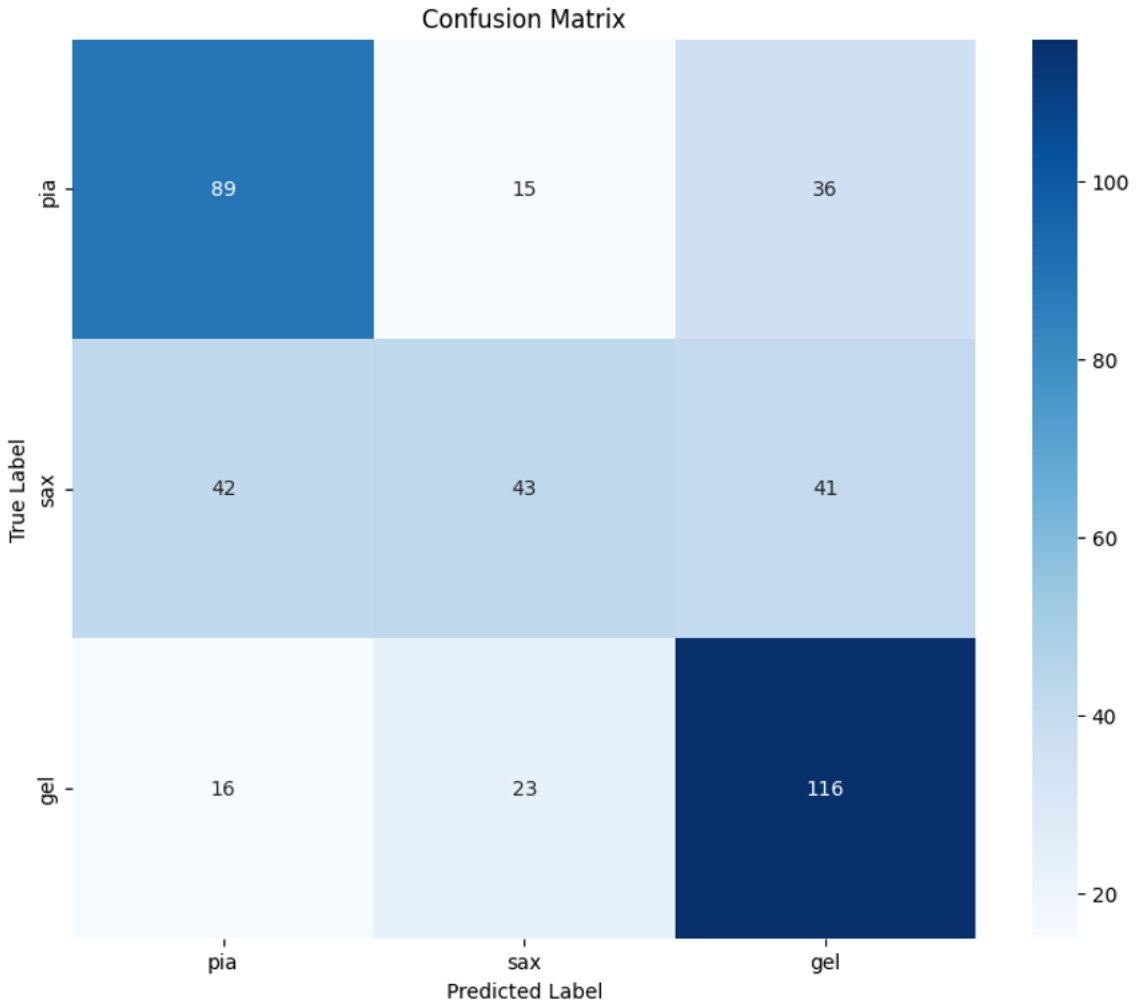
Dal risultato dell'addestramento emerge come i modelli, nonostante la loro semplicità, rispondono bene agli input: in particolare i modelli che utilizzano lo spettrogramma di Mel e il cromogramma raggiungono un'accuratezza rispettivamente di 0.59 e 0.60.

Nel primo caso si ha che il pianoforte ha una precision di 0.63 e una recall di 0.76, la chitarra elettrica una precision di 0.67 e una recall di 0.66 mentre il sassofono una precison di 0.41 e una recall di 0.30.

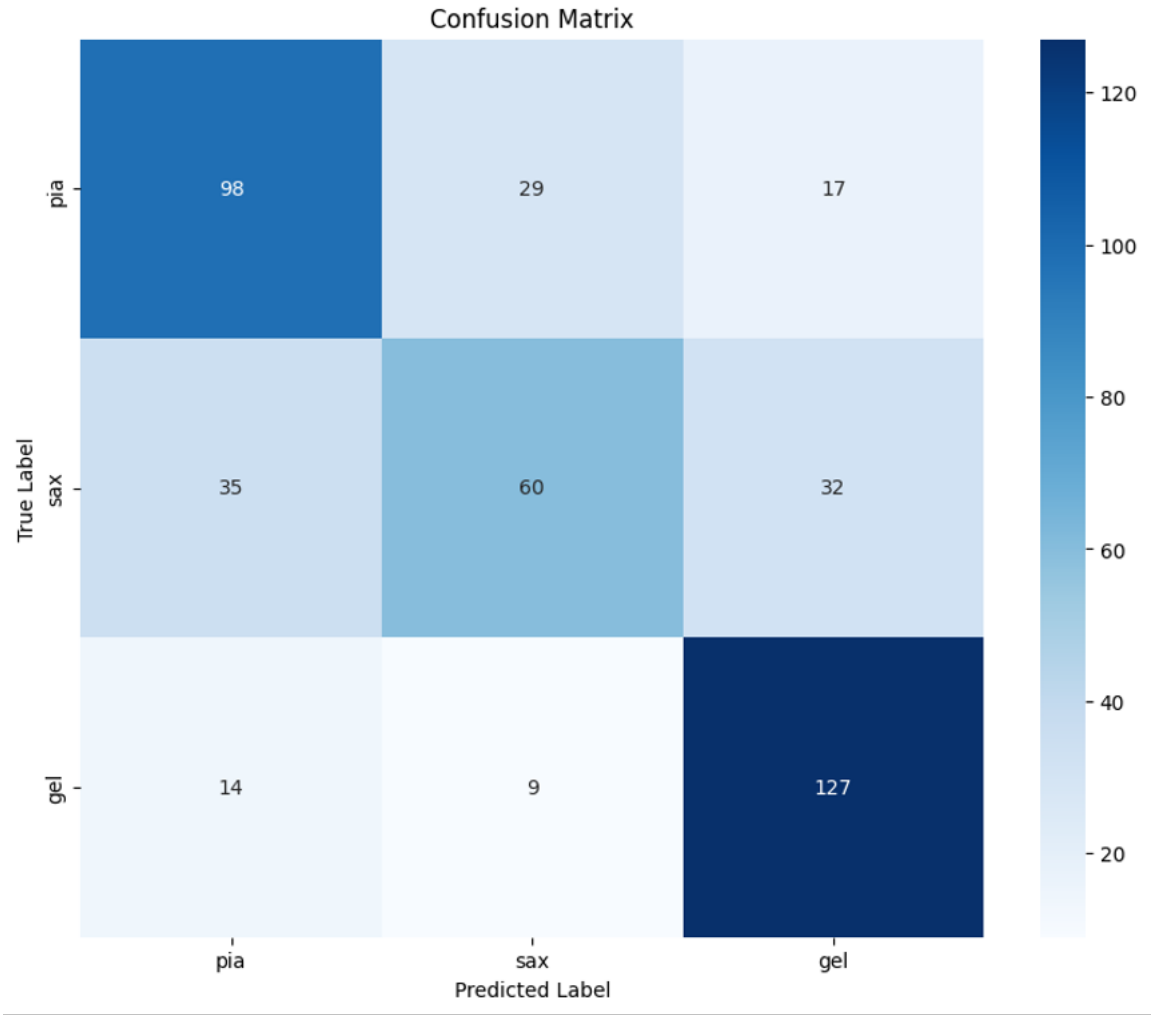


Si hanno prestazioni simili anche nel secondo caso, dove il pianoforte ha una precision di 0.61 e una recall di 0.60, la chitarra elettrica una precision di 0.75 e una recall di 0.66 mentre il sassofono una precison di 0.53 e una recall di 0.34.

Quello che si nota è che la classe che viene maggiormente confusa in entrambi i casi è il sassofono, che non viene quasi mai riconosciuto correttamente.



Utilizzando come input waveplot e spettrogramma CQT l'accuracy migliora notevolmente, raggiungendo un valore di 0.68. In questo caso il pianoforte ha una precision di 0.67 e una recall di 0.68, la chitarra elettrica una precision di 0.72 e una recall di 0.85 mentre il sassofono una precison di 0.61 e una recall di 0.47.



Si nota che la classe più solida è la chitarra elettrica, mentre il sassofono, pur rimanendo la classe più problematica, migliora sensibilmente rispetto ai casi precedenti, in particolare nella precision. Questo indica che il modello ha classificato correttamente una buona parte degli elementi predetti come ‘sassofono’, ma continua a classificare erroneamente molti elementi appartenenti alla classe ‘sassofono’.

Conclusioni e Sviluppi futuri

Dai dati emersi si evince come il miglior spettrogramma per classificare gli strumenti musicali sia quello che utilizza la trasformata CQT, che riesce a catturare le informazioni più determinanti per distinguere uno strumento musicale da un altro.

Eventuali miglioramenti di questo modello si avrebbero ampliando il dataset, così da poter ampliare anche il numero di classi; si potrebbe anche convertire la task da instrument classification a instrument tagging, costruendo sulla base di questo un nuovo modello in grado di riconoscere tutti gli strumenti musicali che concorrono in una singola traccia.