Analisi del parlato: sentimenti, emozioni e dialetto.

L'analisi dell'audio è molto importante nel mondo odierno perché sempre più interazioni sono fatte tramite supporti interattivi vocali. La principale informazione estratta dal vocale è la semantica, in altre parole il significato dei concetti espressi, ma al suo interno sono contenute altrettante informazioni. Alcuni ambiti di applicazione sono il riconoscimento vocale, dei sentimenti, delle emozioni, la provenienza geografica, ma anche informazioni mediche, come per esempio indicazioni sui sintomi di malattie. La risoluzione dei compiti legati all'audio, ed in particolare alla voce, è molto impegnativa e le tecniche usate per risolvere questi problemi si basano spesso su approcci di Deep Learning che richiedono elevate potenze di calcolo.

Introduzione

Le interazioni vocali sono sempre più diffuse perché permettono di migliorare in modo sensibile l'interazione sia tra utenti sia tra utente e software/macchine, ad esempio: i controlli vocali all'interno delle automobili, sugli smartphone, ma anche per applicazioni industriali e, più in generale, tutte quelle situazioni in cui non si possono, o non si vogliono, usare le mani per digitare del testo su una tastiera. Per questo, l'analisi dell'audio è diventato un argomento molto importante nel mondo contemporaneo in quanto, rispetto ad esempio a un file di testo, le informazioni contenute all'interno di una frase parlata sono molto più dettagliate. Ad esempio, nel mondo delle vendite è sempre più difficile affermarsi e rendersi appetibili agli occhi di potenziali clienti e per questo avere informazioni riguardo le conversazioni può giocare un ruolo determinante sia nella fase di vendita sia in fase di analisi. Durante una trattativa è di fatti possibile raccogliere molte informazioni che possono migliorare l'esperienza di vendita: capire velocemente il grado di interesse delle parti per valutare il proseguimento, oppure modulare un'offerta per renderla più appropriata; a livello strategico è importante per i manager conoscere l'andamento delle campagne e poter analizzare dati aggregati, affiancati da opportune metriche di valutazione, da poter esplorare e su cui basare le decisioni future.

In letteratura, problemi del genere solitamente sono risolti con tecniche di deep learning, infatti un grande limite a cui si va incontro quando si lavora con l'audio è senza dubbio la grande mole di dati necessaria all'addestramento di modelli di questo tipo. Inoltre, gran parte dei modelli disponibili sono stati sviluppati per la lingua inglese e applicazioni per la lingua italiana sono meno diffuse e molte volte non sono sufficienti per ottenere delle buone prestazioni; questo comporta che sia spesso necessario passare da una fase di raccolta dati che molte volte può essere lunga e costosa.
Andando più nello specifico si possono identificare quattro task riguardanti l'analisi dell'audio: speech recognition, sentiment analysis, emotion analysis e riconoscimento del dialetto.

Speech recognition

La speech recognition è il cardine dell'analisi dell'audio: a partire da un ingresso audio vocale, lo scopo è trascrivere il suo contenuto in un documento di testo e per questo è quindi equiparabile ad una versione più complessa di un classificatore di suoni. Infatti, molte volte i problemi di questo

tipo sono modellati come classificatori di fonemi, poi eventualmente seguiti da altri modelli per la predizione del testo in base ai fonemi riconosciuti.

Riconoscimento dei dialetti

Senza ombra di dubbio il task su cui si hanno meno difficoltà a comprenderne la natura è l'identificazione di un dialetto. In questo caso, infatti, partendo da un file audio, lo scopo finale è identificare la provenienza dell'interlocutore. Seppure una nazione possa condividere un'unica lingua molte volte tra territori differenti quest'ultima varia dando origine a dei dialetti. Il caso limite, ad esempio, è in Italia: la lingua nazionale è l'Italiano, ma in Sardegna si parla il Sardo che è considerato come una lingua autonoma piuttosto che un dialetto.

Figura 1: Raggruppamento dei dialetti in Italia [7]

In generale i dialetti però tendono ad avere alcune somiglianze tra di loro, quindi se anche questo può essere considerato alla stessa stregua di un problema di classificazione linguistica diventa decisamente più complicato.

Sentiment ed Emotion Analysis

Molte volte si tende ad usare in maniera intercambiabile l'espressione *sentiment analysis* con *emotion analysis.* La sostanziale differenza tra analisi dei sentimenti e analisi delle emozioni risiede nel numero di classi. L'analisi dei sentimenti è riconducibile ad un problema di classificazione binario, in cui l'input può essere assegnato ad una classe *negative* oppure *positive*. In alcuni casi invece si preferisce introdurre un ulteriore livello di precisione inserendo la classe *neutral.*
L'analisi delle emozioni invece è un problema più articolato e difficoltoso da gestire. Al contrario del primo tipo di analisi non ci sono due o tre classi con evidenti differenze, ma in questo caso si hanno più classi che possono avere similarità tra di loro. Ad esempio, si può avere la classe *happiness* ed *excitement* che pur essendo due tipi di emozioni differenti hanno diverse caratteristiche in comune come il volume o il tono della voce.
In alcuni contesti è possibile vedere l'analisi delle emozioni come un'estensione dell'analisi dei sentimenti, il che porta ad ottenere una visione più dettagliata e meno semplificata di una conversazione. Ad esempio, *angry* oppure *bored* va da sé che sono due emozioni che non sono

propriamente positive, infatti in un eventuale analisi dei sentimenti verrebbero inserite nella classe *negative.*
Per via della natura dei problemi talvolta, quando si hanno a disposizione solo informazioni testuali, come possono essere recensioni di un prodotto, si tende a preferire l'analisi dei sentimenti rispetto all'analisi delle emozioni in quanto non avendo audio o immagini il primo task è nettamente più difficoltoso.

Tecniche per analizzare l'audio

L'analisi dell'audio può essere fatta tramite diverse tecniche, ognuna con un tipo di pre-processamento del file audio differente. La più naturale che può venire in mente è sicuramente effettuare il riconoscimento del parlato per poi lavorare sul testo tramite strumenti di Natural Language Processing. Altre tecniche, un po' meno intuitive, passano per la generazione di feature audio che verranno poi analizzate tramite un classificatore realizzato tramite una Rete Neurale.

Classificazione tramite Reti Neurali

Una Rete Neurale artificiale non è altro che un modello matematico il cui scopo è approssimare una determinata funzione. La sua composizione è detta a livelli in quanto contiene un primo livello di input, dei livelli intermedi detti nascosti ed un ultimo livello di output. Ogni livello nascosto della rete è composto da un numero predefinito di unità di elaborazione, dette neuroni. Ogni neurone al livello n ha diversi segnali in ingresso provenienti dai neuroni al livello $n-1$ ed un segnale di uscita per i neuroni al livello $n+1$.

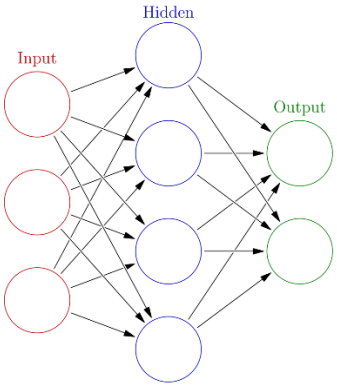


Figura 2: Rete neurale artificiale [2]

Le reti più comuni e che descriveremo di seguito sono dette *feedforward* in quanto il loro grafo di computazione è diretto ed aciclico. Vediamo ora come funziona l'apprendimento da parte di una rete neurale. Bisogna innanzitutto descrivere più nel dettaglio la struttura interna di un neurone. Come precedentemente detto la singola unità computazionale ha un ingresso ed un'uscita, ma il dato come *trasformato* al suo interno?

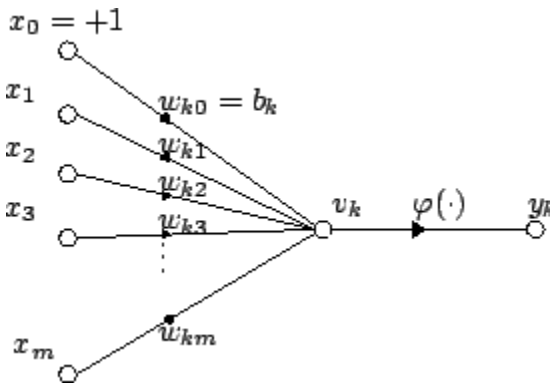


Figura 3: Neurone artificiale [3]

Definiamo per ogni connessione i, entrante in un neurone, un input x_i ed un peso associato w_i; gli input vengono quindi combinati linearmente con i pesi associati effettuando l'operazione $a = w^T x$. Il risultato di questa operazione viene dato come input ad una funzione detta *funzione di attivazione.* Riassumendo e formalizzando in maniera più concisa le operazioni effettuate all'interno di un singolo neurone si ha che l'output è:

$$output = \sigma(a) = \sigma(w^T x) = \sigma\left(\sum_i w_i x_i\right)$$

La *funzione di attivazione* è quel componente del modello che si occupa di determinare il valore di uscita in base alla combinazione lineare dei pesi del neurone con gli output dei layer precedenti. Esistono diverse funzioni di attivazione: allo scopo attuale è sufficiente sapere che per i layer intermedi viene spesso usata la funzione (non lineare) chiamata ReLU; mentre se si risolve un problema di classificazione sull'ultimo layer viene usata una funzione detta Softmax, la cui proprietà è restituire in output una distribuzione di probabilità sulle varie classi.

L'ultimo nodo da sciogliere ora riguarda l'addestramento di un modello di questo tipo.
L'addestramento di una Rete Neurale avviene in due fasi: una prima fase, detta di forward propagation, in cui le si danno in pasto dei dati presi da un insieme di addestramento ed una seconda fase, detta di backpropagation, in cui si fa il vero e proprio addestramento.
Nella forward propagation si utilizza una parte del dataset chiamata *train set* per calcolare la predizione del modello e confrontarlo con la realtà. Il confronto, calcolato per mezzo di una funzione di *loss*, è una misura oggettiva di quanto la predizione del modello sia corretta.
L'addestramento è l'ottimizzazione matematica della funzione di *loss*: in altre parole i pesi del modello sono modificato nel tentativo di ridurre la funzione di *loss* e quindi far sì che la predizione del modello sia il più vicino possibile alla realtà. Si fa uso quindi, nella fase di back propagation, di una regola per il calcolo delle derivate detta *chain rule*; in questo modo, calcolando la derivata della funzione di *loss* rispetto ai pesi dei livelli, possiamo variare i pesi per cercare di diminuire il valore della funzione oggetto di ottimizzazione.

Preprocessing dell'audio

I file audio, prima di essere dati in input ad un classificatore, devono subire una fase di preprocessing dove si vanno ad estrarre dei coefficienti detti feature.
Un file audio ha una caratteristica detta *frequenza di campionamento.* Solitamente si assume essere 16kHz, questo vuol dire che in un secondo di audio ci sono sedicimila campioni.

Assumeremo questa frequenza come uno standard negli esempi successivi. Calcolare un set di coefficienti per ogni campione è un'operazione molto onerosa, oltre che inutile in quanto la variazione tra un campione ed il suo successivo, o precedente, è minima. Si considerano quindi batch di campioni consecutivi di dimensione 400, corrispondenti quindi a 25ms. Per ogni insieme di campioni si andranno a calcolare le feature audio. Per i task di speech recognition in genere sono usate un tipo particolari di feature chiamate *Mel-Frequency Cepstral Coefficent (MFCC)*; che si calcolano con il seguente procedimento:

1. Preso il batch di frame viene calcolata la Trasformata di Fourier. Considerando:
   a. $x_i(n)$ batch di segnali campionati con $0 \leq n \leq 399$
   b. si calcola $X_i(k) = \sum_{n=0}^{N=399} x_i(n) w(n) e^{-j\frac{2\pi kn}{N}}$ con $0 \leq k \leq N$ e $w(n)$ la hamming window.

   E data la frequenza di campionamento
   $$f_s = 16000$$
   si avrà che la frequenza corrispondente a k è data da
   $$l_f(k) = \frac{k f_s}{N}.$$
   Calcolandolo per tutti i batch di campioni otterremo quindi
   $$X = [X_1, X_2, \dots, X_p]$$
   detta *Short Time Fourier Transform Matrix (STFT Matrix)*
2. Si effettua il calcolo del Mel Filterbank:
   a. Tramite formula di conversione si passa dalla scala delle frequenze alla scala Mel:
   $$\Phi_f = 2595 * log_{10}\left(1 + \frac{l_f}{700}\right)$$
   b. Dati:
   $$l_{f_{max}} \text{e } l_{f_{min}} \text{si calcola } \Phi_{f_{max}}, \Phi_{f_{min}} e\ \delta\Phi_f = \frac{\Phi_{f_{max}} - \Phi_{f_{min}}}{F+1}$$
   Con F numero di filtri.
   c. Si calcola
   $$\forall\ 1 \leq m \leq F, \Phi_{f_c}(m) = m.\delta\Phi$$
   dopodiché, tramite formula inversa, si porta questo valore in scala delle frequenze ottenendo
   $$l_{f_c}(m)$$
   d. Il banco di filtri Mel è dato da questa funzione:
   $$M(m,k) = \begin{cases} 0 & \text{for } l_f(k) < l_{f_c}(m-1) \\ \frac{l_f(k) - l_{f_c}(m-1)}{l_{f_c}(m) - l_{f_c}(m-1)} & \text{for } l_{f_c}(m-1) \leq l_f(k) < l_{f_c}(m) \\ \frac{l_f(k) - l_{f_c}(m+1)}{l_{f_c}(m) - l_{f_c}(m+1)} & \text{for } l_{f_c}(m) \leq l_f(k) < l_{f_c}(m+1) \\ 0 & \text{for } l_f(k) \geq l_{f_c}(m+1) \end{cases}$$
3. Si procede al calcolo di MFCC:
   a. Si passa alla scala logaritmica delle Mel:
   $$L_p(m,k) = ln\left(\sum_{k=0}^{N-1} M(m,k)\left|X_p(k)\right|\right)$$
   b. Si applica la *discrete cosine transform*:
   $$\phi_p^r\big(x(n)\big) = \sum_{m=1}^{F} L_p(m,k) cos\left(\frac{r(2m-1)}{2F}\right)$$
   ottenendo così che la colonna p-esima della matrice Φ rappresenta i coefficienti MFCC del segnale corrispondente.

Da calcolo di questi coefficienti si può generare spettrogrammi o matrici con cui poi si andranno a effettuare le relative analisi.

Analisi dell'audio tramite Reti Ricorrenti

Al giorno d'oggi due tipi di reti ricorrenti molto efficaci sono le Long Short Term Memory (LSTM) e le Bidirectional LSTM (BiLSTM). Per una descrizione accurata su questo tipo di struttura si rimanda ad un articolo precedentemente pubblicato[1].
Il grande svantaggio di questo tipo di strutture però è l'impossibilità di variare la dimensione dell'input in quanto ricevono sempre un vettore di dimensione fissata.
Tuttavia, esistono dei lavori recenti basati sul *meccanismo dell'attenzione* che risolvono questo problema e al contempo riescono a gestire in maniera più efficace le dipendenze a lungo termine.
A titolo esemplificativo è possibile descrivere uno dei più semplici meccanismi dell'attenzione come un vettore. Definiamo innanzitutto

$$H = [h_1, h_2, \dots, h_T]$$

come la matrice di input di forma T x F, dove T è il numero di frame e F è il numero di feature per frame. Il meccanismo di attenzione è un vettore

$$\omega \in R^F$$

tale per cui, se moltiplicato per H, formando

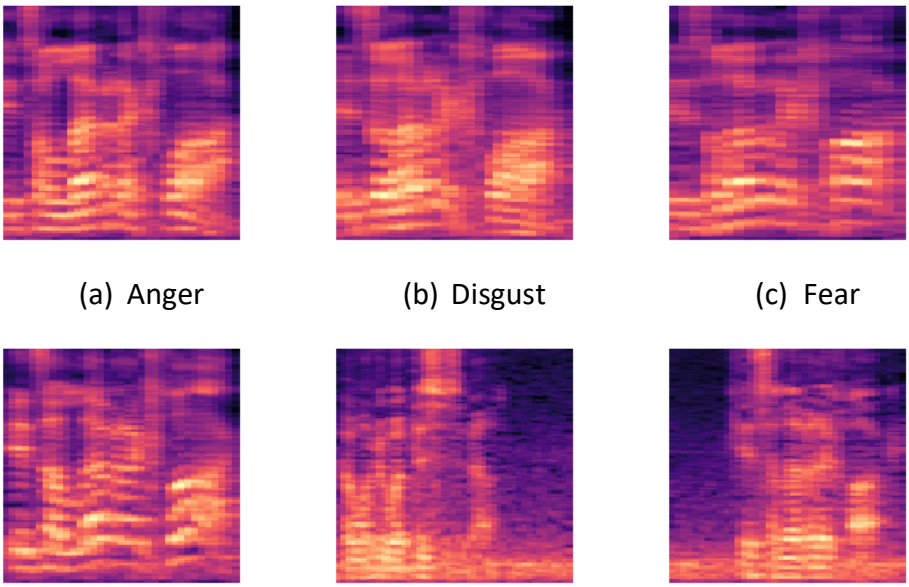
$$M = \omega^T H$$

produce una *attention map*. Un utilizzo possibile per questa *attention map* può essere come fattore *pesante* per l'output di una RNN, in maniera tale da dire al modello *dove prestare più attenzione*.

Analisi dell'audio tramite CNN

Dalle feature audio estratte durante il preprocessing è possibile generare uno spettrogramma, ovvero una rappresentazione grafica tridimensionale sottoforma di immagine che mostra variazione dello spettro delle frequenze secondo il tempo. Sull'asse orizzontale c'è quindi il tempo, mentre sull'asse verticale è presente la frequenza. La terza dimensione è rappresentata dalla scala di colori utilizzata per mostrare la variazione di ampiezza di una frequenza.

[1] https://www.industry4business.it/servitization/predictive-maintenance/deep-learning-con-lstm-per-la-manutenzione-predittiva/



(a) Anger (b) Disgust (c) Fear

(d) Happiness (e) Sadness (f) Neutral

Figura 4: Esempi di spettrogrammi per ciascuna classe di emozione.

A tutti gli effetti, essendo quindi immagini, ci si riconduce ad un problema di classificazione di immagini e quindi diventa possibile usare le CNN, o modelli basati su di essa, per classificarli. Per una rapida e dettagliata spiegazione di cosa sia una Convolutional Neural Network si rimanda ad un articolo precedentemente pubblicato[2].
Al giorno d'oggi esistono molti modelli di Deep Learning adatti ad effettuare Image Classification, uno dei più famosi è VGG in tutte le sue possibili declinazioni.
Inoltre, grazie a preaddestramenti effettuati su altri dataset è possibile raggiungere risultati soddisfacenti con relativamente pochi dati a nostra disposizione. Infatti, molte volte nelle librerie di DL che implementano questi modelli si può scegliere anche se scaricare un set di pesi da cui far partire l'addestramento. In questo caso si parla di Transfer Learning o Fine Tuning.

Conclusioni

L'analisi dell'audio si sta diffondendo in molti campi di applicazione e permette attraverso l'interazione vocale di comandare software e macchine risolvendo molte situazioni in cui l'utente è impossibilitato, o ostacolato, nell'interazione fisica con il device. Attraverso strumenti di deep learning è possibile estrarre molte informazioni dall'interazione vocale, ad esempio riguardanti l'emozione o la provenienza dell'interlocutore, e questo permette di migliorare l'interazione e l'esperienza dell'utente, ma porta anche a raccogliere nuovi dati su cui generare degli indicatori di performance molto più esaustivi e precisi.

L'utilizzo di questi strumenti sta avendo un impatto molto positivo nell'ambito dell'assistenza telefonica in quanto permette una migliore gestione delle chiamate, un'accurata analisi in tempo reale e la pianificazione di strategie più efficaci. Inoltre, è possibile usare dei modelli di *question answering* per supportare le risposte in maniera molto più rapida dei classici strumenti di ricerca. Infine, in ambito di prevenzione medica, le ultime ricerche hanno investigato la possibilità di usare l'analisi dell'audio, in particolar modo dei colpi di tosse, per potere aiutare i medici

[2] https://www.agrifood.tech/analisti-ed-esperti/smart-agriculture-deep-learning/

nell'individuazione di soggetti positivi alla COVID-19[3]. Gli studi [8], seppur preliminari, sembrano molto promettenti e permetterebbero di effettuare degli screening veloci, non invasivi e su larga scala con semplici strumenti di raccolta audio come gli smartphone.

In sintesi, i pregi e i difetti dell'utilizzo di tecniche di analisi audio sono:

- Pro:
  - Acquisizione di nuovi dati e realizzazione di KPI più efficaci
  - Ottimizzazione del flusso di lavoro
  - Miglioramento delle interazioni uomo-macchina
- Contro:
  - Necessità di molti dati con annotazioni
  - Difficoltà di reperimento di dataset italiani
  - Gestione della privacy e del contenuto dei dati

Riferimenti

[1] Choi, Keunwoo, et al. "A comparison of audio signal preprocessing methods for deep neural networks on music tagging." *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018.

[2] https://en.wikipedia.org/wiki/Artificial_neural_network#/media/File:Colored_neural_network.svg

[3] https://en.wikipedia.org/wiki/Artificial_neuron

[4] Muaidi, Hasan, et al. "Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques." *Research Journal of Applied Sciences, Engineering and Technology* 7.24 (2014): 5082-5097.

[5] Kopparapu, Sunil Kumar, and M. Laxminarayana. "Choice of Mel filter bank in computing MFCC of a resampled speech." *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*. IEEE, 2010.

[6] Ramet, Gaetan, et al. "Context-aware attention mechanism for speech emotion recognition." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

[7] https://it.wikipedia.org/wiki/Lingue_dell%27Italia

[8] *Laguarta et al. "COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings", IEEE Open Journal of Engineering in Medicine and Biology, IEEE, 2020.*

[3] https://news.mit.edu/2020/covid-19-cough-cellphone-detection-1029