**Documentazione di Progetto del Corso Ingegneria della Conoscenza**

A.A. 2020/2021

**Automatic Speech Recognition (ASR) attraverso la creazione, training ed inferenza su un Recurrent Neural Network per la generazione automatica di sottotitoli**

Di:

|  |  |
| --- | --- |
| Luigi Porcelli  Matricola: 701719 | Nicolò Cucinotta  Matricola: |

**Obiettivo:**

L’obiettivo del nostro lavoro è stato quello di costruire un modello di Recurrent Neural Network e addestrarlo su un dataset di composto da file audio e relative trascrizioni al fine da ottenere un modello capace analizzare un audio e generare la trascrizione del suddetto.

Successivamente, il nostro obiettivo è stato quello di poter effettuare inferenza sul nostro modello al fine di poter generare automaticamente dei sottotitoli per un dato video.

**Strumenti utilizzati:**

Il progetto è stato sviluppato in linguaggio Python con l’utilizzo di Tensorflow, framework open-source della Google per il machine learning.

Le librerie utilizzate sono state:

* Keras per
* Librosa per

Per il training, è stato utilizzato Google Colab con accelerazione hardware TPU. Per poter usufruire della TPU abbiamo inoltre utilizzato Google Cloud Storage, per necessità di Colab stesse.

**Dataset e Preprocessing:**

Il dataset da noi utilizzato per il training è stato [LibriSpeech](https://www.openslr.org/12), in particolare il dataset train-clean-360, composto da (tanti) file audio di parlato inglese per un totale di 360 ore di parlato. Il dataset è derivato da audiolibri.   
  
Tale dataset ha subito una fase di preprocessing: è stato necessario processare, grazie a funzioni di Librosa, sia i file audio che le trascrizioni.

Gli *audio* sono stati processati al seguente modo:   
in base alle nostre ricerche, la rappresentazione migliore dei dati per il nostro task erano i [Mel Spectrogram](https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53), ossia spettrogrammi in scala Mel. Una volta ottenuti tali spettrogrammi, abbiamo calcolato il logaritmo su di essi per (some-reasons)

Le *trascrizioni* sono state processate al seguente modo:   
Ci siamo serviti di un vocabolario composto da tutte le lettere dell’alfabeto con l’aggiunta dell’apostrofo (‘) e dello spazio ( ), con l’aggiunta di un carattere vuoto (“”) per indicare un momento di silenzio o un carattere non identificato dal modello. In totale, quindi, il vocabolario è composto da 29 segni. Il passo successivo è stato quello di mappare ogni elemento del vocabolario ad un intero da 0 a 28: le trascrizioni erano così una sequenza ordinata di interi nel range 0-28. La mappatura è stata necessaria per la dimensione di output della rete neurale.

Infine, le coppie audio-trascrizione processate sono state salvate in file di formato TFRecord, per migliorare le performance in fase di training su Google Colab.

**Modello e training:**

Il modello è un modello di RNN, adatto per gli audio

Il training è stato effettuato sul

**Inferenza:**

Il progetto è stato sviluppato in linguaggio Python con l’utilizzo di Tensorflow, framework open-source della Google per il machine learning.