RELAZIONE:
Text-Summarizer

Enrico Ferraiolo 0001191698

Laurea Magistrale in Informatica

Corso: Natural Language Processing
a.a. 2024-2025

Indice

1 Introduzione

Questo progetto vuole implementare un modello di text summarization (riassunto dei testi) utilizzando un'architettura Sequence-to-Sequence (Seq2Seq) basata su reti LSTM (Long Short-Term Memory).
L'obiettivo principale è generare riassunti concisi e significativi a partire da recensioni di prodotti più lunghe, mantenendo il significato del testo originale.

2 Dataset

Per questo progetto è stato utilizzato il dataset SNAP Amazon Fine Food Reviews, che contiene recensioni di prodotti alimentari di Amazon.
In particolare, il dataset contiene, per ogni riga, una recensione completa e il rispettivo riassunto.
Del dataset originale, composto da circa 500.000 righe, è stato selezionato un sottoinsieme di 10.000 righe per l'analisi e l'allenamento del modello.

3 Preprocessing dei Dati

Il preprocessing dei dati è una fase critica per garantire la qualità e l'efficacia del modello di summarization, infatti è fondamentale pulire e filtrare i dati in modo accurato.
Sul dataset, infatti, sono stati eseguiti diversi passaggi di pulizia e filtraggio dei dati per garantire qualità e coerenza al modello durante l'addestramento.
Vediamo di seguito gli step effettutati durante questa fase:

3.1 Pulizia del Testo

Sono stati applicati i seguenti step di preprocessing:

1. **Conversione del testo in minuscolo**
   - Questa conversione garantisce l'uniformità del testo, evitando che la stessa parola venga considerata diversa solo per la presenza di maiuscole. Ad esempio, "Home", "HOME" e "home" vengono trattate come la stessa parola, riducendo la dimensionalità del vocabolario e migliorando l'efficienza dell'addestramento.
2. **Rimozione dei tag HTML**
   - Le recensioni potrebbero contenere tag HTML residui dal formato web originale. Questi elementi non contribuiscono al significato semantico del testo e potrebbero interferire con l'apprendimento del modello, pertanto vengono rimossi.
3. **Espansione delle contrazioni**
   - Le contrazioni nella lingua inglese (come "don't", "I'm", "we're") vengono espanse nelle loro forme complete ("do not", "I am", "we are"). Questo processo vuole standardizzare e garantire coerenza in tutto il testo e aiuta il modello a catturare meglio le relazioni semantiche, eliminando variazioni non necessarie della stessa espressione.

4. **Rimozione degli apostrofi possessivi ('s)**

    - La forma possessiva in inglese non altera sostanzialmente il significato della frase ai fini del riassunto.
      La sua rimozione semplifica il testo e riduce ulteriormente la dimensione del vocabolario, permettendo al modello di concentrarsi sui concetti principali.

5. **Eliminazione del testo tra parentesi**

    - Il testo tra parentesi spesso contiene informazioni supplementari che non sono generalmente essenziali per il riassunto.
      La loro rimozione aiuta a mantenere il focus sulle informazioni principali della recensione.

6. **Rimozione della punteggiatura e caratteri speciali**

    - La punteggiatura e i caratteri speciali, pur essendo importanti per la leggibilità umana, possono introdurre rumore nell'addestramento del modello.
      La loro rimozione semplifica il testo mantenendo intatto il contenuto semantico essenziale per la generazione del riassunto.

7. **Eliminazione delle stopwords**

    - Le stopwords sono parole molto comuni (come "the", "is", "at", "which") che appaiono frequentemente ma portano poco significato semantico.
      La loro rimozione riduce significativamente la dimensionalità del problema senza perdere informazioni cruciali per il riassunto, permettendo al modello di concentrarsi sulle parole più significative.

8. **Rimozione delle parole troppo corte**

    - Le parole molto corte (solitamente di una o due lettere) spesso non contribuiscono al significato del testo.
      La loro rimozione aiuta a ridurre ulteriormente il rumore nei dati, mantenendo solo i termini più significativi per l'analisi.

3.2 Filtraggio dei Dati

Dopo l'analisi statistica del dataset, sono stati applicati i seguenti vincoli: TODO: INSERISCI IMMAGINE(?)

- Lunghezza massima delle recensioni: 30 parole
- Lunghezza massima dei riassunti: 8 parole

Questi limiti sono stati determinati attraverso un'analisi statistica della distribuzione delle lunghezze nel dataset.

3.3 Tokenizzazione e Token Speciali

Per preparare i dati per il modello:

- Sono stati aggiunti token speciali:

  - "sostok" come marcatore di inizio sequenza
  - "eostok" come marcatore di fine sequenza
- È stata effettuata la tokenizzazione separata per:
  - Recensioni (testo di input)
  - Riassunti (testo di output)

4 Architettura del Modello

L'implementazione del modello è stata strutturata seguendo i principi della programmazione orientata agli oggetti, con una chiara separazione delle responsabilità.

4.1 Classe Base Astratta

La classe `BaseModel` fornisce l'interfaccia base per tutti i modelli di summarization:

- Definisce i metodi astratti per la costruzione dell'encoder e del decoder
- Implementa funzionalità comuni come il salvataggio del modello e l'inferenza
- Gestisce la conversione tra sequenze e testo

4.2 Implementazione Seq2SeqLSTM

La classe `Seq2SeqLSTM` implementa l'architettura specifica:

4.2.1 Encoder

L'encoder è composto da:

- Layer di embedding con dimensione 100
- Tre layer LSTM in cascata con:
  - Dimensione latente di 300 unità
  - Dropout del 40% per regolarizzazione
  - Return sequences e return state attivati

4.2.2 Decoder

Il decoder include:

- Layer di embedding dedicato
- Layer LSTM con:
  - Stessa dimensione latente dell'encoder
  - Dropout del 40%
  - Recurrent dropout del 20%
- Meccanismo di attention per focalizzarsi sulle parti rilevanti dell'input
- Dense layer con softmax per la generazione del vocabolario di output

5 Training del Modello

Il training è stato eseguito utilizzando:

- Split del dataset in training e validation set
- Monitoraggio della loss e validation loss durante l'addestramento
- Early stopping per prevenire l'overfitting

6 Metriche di Valutazione

Per valutare le performance del modello sono state utilizzate diverse metriche:

6.1 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Sono state calcolate tre varianti di ROUGE:

- ROUGE-1: confronta unigrammi tra il riassunto generato e quello di riferimento
- ROUGE-2: considera bigrammi per catturare la fluidità del testo
- ROUGE-L: identifica la sottosequenza comune più lunga

6.2 Word Error Rate (WER)

Il WER misura:

- Il numero di operazioni necessarie per trasformare il testo generato in quello di riferimento
- Include inserzioni, cancellazioni e sostituzioni di parole

6.3 Cosine Similarity

Questa metrica:

- Valuta la similarità semantica tra i vettori dei testi
- Fornisce un valore tra -1 e 1, dove 1 indica massima similarità

7 Conclusioni

Il modello implementato dimostra la capacità di generare riassunti efficaci delle recensioni di prodotti. L'architettura Seq2Seq con attention e la struttura modulare del codice permettono:

- Facile estensibilità per futuri miglioramenti
- Buone performance nella generazione di riassunti
- Robustezza grazie alle tecniche di regolarizzazione implementate