Alma Mater Studiorum
Università di Bologna
Campus di Cesena

DIPARTIMENTO DI INFORMATICA – SCIENZA E INGEGNERIA
Corso di Laurea in Ingegneria e Scienze Informatiche

TRADUZIONI AUTOMATICHE

Relatore:
Prof. Giovanni Delnevo

Presentata da:
Alessandro Pioggia

Sessione III
Anno Accademico 2021-2022

Desidero dedicare la tesi a coloro che lo hanno reso possibile.

Introduzione

Questa è l'introduzione.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Contesto

1.1 Neural machine translation - NMT

La neural machine translation (in italiano traduzione automatica neurale (NMT)) è un approccio alla traduzione automatica che utilizza una rete neurale per prevedere la probabilità di una sequenza di parole, modellando in genere intere frasi in un unico modello integrato. La creazione del modello sfrutta interamente il machine learning, dunque è in grado di apprendere autonomamente sulla base dei training alla quale prende parte.

Differenze dai modelli tradizionali

- Rispetto ai modelli di traduzione tradizionali, chiamati SMT(statical machine translation), sfruttano una quantità infinitesimale di memoria;
- grazie all'utilizzo del deep learning nella creazione della rete neurale, è possibile optare per un addestramento del tipo end-to-end, questo permette di semplificare la pipeline di sviluppo, con conseguente aumento di prestazioni non indifferente. Per andare più nello specifico:
  - una pipeline tradizionale ha questo aspetto:
    voce(input) → riduzione del rumore (noise) → estrazione dei fonemi → composizione delle parole → trascrizione

– una pipeline che sfrutta il deep learning si presenta invece così: voce(input) $\rightarrow DNN \rightarrow$ trascrizione

Osservando le pipeline proposte, è possibile osservare che l'approccio tradizionale porta con sè una complessità non indifferente, ogni singolo layer citato deve essere ottimizzato separatamente, attraverso un preciso criterio. La pipeline creata a partire da una rete neurale invece, semplifica la comunicazione fra i due agenti(input $\rightarrow$ output desiderato).
Nota: DNN = deep neural network

1.1.1 RNN

Prima di descrivere in maniera analitica e precisa la struttura che permette la costruzione di un modello di encoding-decoding è necessario definire le RNN, o Recurrent Neural Networks. Le RNN sono delle reti neurali molto interessanti, perché a differenza dalle tradizionali straight forward neural networks, in cui l'informazione può andare solo in un verso (in avanti), in questo tipo di reti i nodi possono essere interconnessi ai livelli precedenti, ammettendo dei loop. Il fatto che i neuroni (o nodi, intesi come elementi che compongono la rete neurale) possano propagare il segnale in tutte le direzioni, permette di introdurre intrinsecamente il concetto di memoria. Questo perché, l'output di un neurone può ipoteticamente influenzare sè stesso in un diverso istante temporale (rispetto al presente). In particolare, una RNN richiede, oltre all'input tradizionale, lo hidden state, che non è altro che l'output prodotto dallo stesso neurone al passo precedente.

Perché sfruttare le RNN? Le RNN, grazie alla memoria, possono operare su dati dinamici mentre le SNN (straight forward neural networks) operano unicamente su dati statici. Grazie a questa proprietà, le reti neurali ricorrenti, sono in grado di trattare delle sequenze temporali, variabili nel tempo. Un esempio di applicazione può convenire nella interpretazione del linguaggio dei segni. In questo ambito applicativo non è richiesto solo di ri-

conoscere la mano, bensì di comprendere la sequenza di movimenti per poter dedurre ed interpretare il significato dei gesti. Quest'ultima constatazione ci rende in grado di percepire l'importanza di questo tipo di rete neurale nell'ambito delle traduzioni automatiche, in quanto si ha a che fare con sequenze di parole, non necessariamente definite in maniera statica che, per dare un senso logico alla frase, necessitano un meccanismo che sfrutti la memoria (e non solo). Per concludere l'approfondimento, considero fondamentale precisare che, by default, la memoria di un neurone è piuttosto breve, dunque gli output generati dai primi layer, difficilmente condizioneranno quelli presenti in livelli più avanzati. A questo proposito sono state studiate delle reti neurali a lunga memoria, quali le LSTM (Long Short Term Memory) e le GRU (Gated Recurrent Units).

1.1.2 Transformers

Nell'ambito del NMT, il modello maggiormente utilizzato è il transformer, che negli ultimi anni ha preso il posto delle reti neurali RNN, descritte nella sezione precedente.

Caratteristiche e peculiarità

La tipologia di modelli citati in questa sezione sono nati con l'obiettivo di evitare la ricorsione, facendo spazio a delle tecnologie che garantissero un tipo di computazione che ammettesse il parallelismo. Le caratteristiche principali sono le seguenti:

- data una sequenza di dati in input, a differenza delle reti RNN, che operano word-by-word (ossia hanno la necessità di analizzare ogni singolo elemento della struttura), il transformers sono in grado di processare l'input **"as a whole"**, ossia come un unico grande dato;
- l'introduzione della **self-attention**, che consente un notevole miglioramento nella predittività, ha dato il nome all'articolo che ha presentato questa tipologia di modelli, ovvero "Attention is all you need";

- il **positional embedding**: si tratta di una tecnica che, insieme all'attention, ha come obiettivo quello di eliminare i processi che sfruttano la ricorsione. In particolare, l'idea è di assegnare dei pesi legati alla posizione di una parola all'interno della frase, in modo da verificare quali siano le tuple di parole che hanno più probabilità di essere accostate.

Il primo punto è fondamentale perché consente al modello di non dipendere da neuroni di layer lontani, processando l'intero input in una sola volta, viene eliminato il concetto di perdita di memoria. Il tutto può essere chiarito attraverso un semplice esempio:
Consideriamo un ipotetico articolo, supponendo che i puntini rappresentino il testo di intermezzo:

La nazionale italiana di basket ..... Pozzecco ha deciso di intervenire in sala stampa, ringraziando lo staff.

Dato questo input, il transformer, analizza l'articolo nel suo intero, mentre un modello tradizionale considera parola per parola, ovvero: {0:La, 1:nazionale, 2: italiana, 3: di, 4: Basket, ..., 1200:Pozzecco }. A giudicare da questo esempio si può osservare che è impossibile mantenere il contesto analizzando una parola per iterazione, specialmente se si ha necessità trovare un nesso fra due parole con indici molto distanti.

Descrizione tecnica dettagliata del modello

Si tratta di un sistema di encoding-decoding. In linea generale, l'encoder mappa l'input ricevuto in un vettore continuo, che contiene tutte le informazioni apprese dai dati ricevuti in ingresso. Il decoder a partire dal vettore continuo, passo per passo, genera un unico output. Il singolo output generato è al contempo condizionato dagli ulteriori vettori continui dati in pasto al decoder in precedenza. Questo passaggio avviene ricorsivamente, fino a quando il decoder non incontra il token <end>, rappresentante la end of sentence (in italiano, fine della frase).

Input embedding

Il primo passo è passare il nostro input al word embedded layer, il quale può essere considerato come una lookup table che contiene dei fattori di rappresentazione per ogni parola. In particolare, ogni parola viene mappata in un vettore avente valori continui, che la rappresentano, dal momento che il sistema apprende attraverso pattern numerici.

Positional encoding

In seguito alla definizione dell'embedded layer è necessario passargli delle informazioni circa la collocazione di ogni vettore all'interno della rete neurale, dal momento che non è più possibile sfruttare la ricorsione a questo scopo. Per fare ciò è stata studiata una brillante soluzione, che verte sull'utilizzo di queste due formule:

- $P(pos, 2i+1) = \cos(\frac{pos}{10.000^{2i/dmodel}})$
- $P(pos, 2i) = \sin(\frac{pos}{10.000^{2i/dmodel}})$

Per ogni step di esecuzione di indice pari, genera un vettore di valori continui attraverso la funzione coseno. Per quelli di indice dispari invece, il vettore viene generato sfruttando il seno. Ad ogni vettore risultante viene sommato l'embedding vector corrispondente (ottenuto al primo step). Per la generazione si utilizzano le funzioni seno e coseno per via delle loro proprietà lineari, che l'intelligenza artificiale è in grado di apprendere con maggiore facilità.

Encoding layer

Come introdotto inizialmente, l'encoder mappa l'input ricevuto in un vettore numerico contenente valori continui, rappresentanti le informazioni apprese dai dati ricevuti in ingresso. Contiene due sottomoduli:

- il primo è una unità dedicata alla self- attention chiamata **multi head attention layer**. Andando nello specifico, ;

- l'altro è una vera e propria **rete neurale feed forward.**

I due layer sono collegati fra loro non direttamente, fra loro si interpone un layer intermedio che si occupa della normalizzazione dell'output.

Multi-head attention layer

L'unità citata nel titolo, implementa uno specifico meccanismo di attention chiamato self-attention. Questa tecnologia consente ad un modello, attraverso principi combinatori accurati, di associare le parole presenti nell'input. Ad esempio, supponendo di avere la frase "ciao", "come", "stai", l'unità potrebbe ipoteticamente correlare : "ciao", "come" oppure "stai", "ciao". Attraverso questo tipo di associazioni, il modello può ad esempio interpretare la tipologia di frase, rendendosi conto in questo caso, di avere a che fare con una domanda. Queste intuizioni, permettono alla rete neurale di strutturare un possibile output.

Perché il meccanismo funzioni l'input deve essere passato come parametro a tre distinti layer, fra loro connessi, i quali genereranno i seguenti vettori continui:

- query vector;
- key vector;
- value vector.

In particolare il query vector contiene la richiesta, generalmente in formato json, che verrà etichettata dal key vector. Questa operazione è molto simile a quella che viene effettuata durante una richiesta ad un motore di ricerca, come ad esempio quello di youtube. Quando viene effettuata la ricerca, google etichetta il testo inserito dall'utente, in particolare associa la ricerca a descrizione, autore o titolo del video. Finita l'etichettatura, interroga il database, cercando i migliori match, ovvero i video strettamente compatibili con la richiesta.

Il vettore delle query dunque, viene moltiplicato per il vettore delle chiavi (o key vector), ottenendo come risultato la matrice degli scores, o punteggi. Le celle della matrice risultante contengono un valore numerico, il quale indica quanto forte sia il collegamento fra una parola e l'altra, ovvero l'attenzione che occorre porre. Quindi, semplificando il concetto, all'interno di un arco temporale ogni parola avrà a disposizione un punteggio in relazione ad ogni altra parola presente. Il focus aumenta con l'aumentare del valore contenuto nella cella.

81	18	91
509	1000	230
300	200	3300

Tabella 1.1: Ipotetica matrice degli scores, per semplicità utilizzerò la sua nomenclatura originale, ovvero score matrix.

Osservando la matrice in formato tabellare sovrastante, si nota subito che l'ordine di grandezza del contenuto delle celle necessita una sorta di normalizzazione, per evitare che, la moltiplicazione con altre matrici, generi valori troppo grandi. A questo proposito si effettuano le seguenti operazioni:

- $\frac{scorematrix}{\sqrt{d_k}} \rightarrow$ divido la matrice per la dimensione delle query e delle keys;
- $softmax(x)_i = \frac{exp(x_i)}{\sum_j exp(x_j)} \rightarrow$ in questo passaggio, viene presa come input la matrice ottenuta al passo precedente, x un elemento arbitrario. La matrice risultante chiamata sealed matrix, conterrà per ogni cella un valore probabilistico, il quale range va da 0 ad 1. Alle probabilità più alte verrà attribuita una maggiore attenzione (da qui, il termine attention), mentre quelle inferiori alla soglia verranno ignorate, perché non utili. This allows the model to be more confident on which words to attend to;

- $sealedmatrix \cdot valuematrix = output \rightarrow$ moltiplico la matrice ottenuta al passaggio precedente, che può essere considerata una matrice di attention, per il vettore dei valori definito inizialmente. L'output, come già accennato, oscurerà le parole con un punteggio softmax più basso, dando la priorità alle altre.

- infine nell'ultimo passaggio, l'*output* ottenuto al passaggio precedente viene passato in input ad un layer lineare, che sarà in grado di processarlo.

L'unità prende il nome di multi-head attention layer, perché i tre vettori di input possono essere scissi prima di iniziare il processo, in questo modo sono in grado di operare in parallelo.

1.2 Seconda Sezione

Ora vediamo un elenco puntato:

- primo oggetto

- secondo oggetto

1.3 Altra Sezione

Vediamo un elenco descrittivo:

OGGETTO1 prima descrizione;

OGGETTO2 seconda descrizione;

OGGETTO3 terza descrizione.

1.3.1 Altra SottoSezione

SottoSottoSezione

Questa sottosottosezione non viene numerata, ma è solo scritta in grassetto.

1.4 Altra Sezione

Vediamo la creazione di una tabella; la tabella 1.2 (richiamo il nome della tabella utilizzando la label che ho messo sotto): la facciamo di tre righe e tre colonne, la prima colonna "incolonnata" a destra (r) e le altre centrate (c):

$(1,1)$	$(1,2)$	$(1,3)$
$(2,1)$	$(2,2)$	$(2,3)$
$(3,1)$	$(3,2)$	$(3,3)$

Tabella 1.2: legenda tabella

1.5 Altra Sezione

1.5.1 Listati dei programmi

Primo Listato

```
        In questo ambiente     posso scrivere      come voglio,
lasciare gli spazi che voglio e non % commentare quando voglio
e ci sarà scritto tutto.
Quando lo uso è meglio che disattivi il Wrap del WinEdt
```

Capitolo 2

Secondo capitolo

Questo è il secondo capitolo.

2.1 Prima Sezione

Questa è la prima sezione.

2.2 Seconda Sezione

Questa è la seconda sezione.

Capitolo 3

Terzo capitolo

Questo è il terzo capitolo.

3.1 Prima Sezione

Questa è la prima sezione.

3.2 Seconda Sezione

Questa è la seconda sezione.

3.3 Terza Sezione

Questa è la terza sezione.

Conclusioni

Queste sono le conclusioni.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque a magna quis nunc venenatis vestibulum. Curabitur commodo efficitur ipsum, non ullamcorper tellus. Duis dictum commodo nisi nec venenatis. Donec euismod pulvinar finibus. Suspendisse lorem mi, suscipit quis faucibus ut, luctus in justo. Cras pulvinar arcu ut ullamcorper pulvinar. Aliquam dictum tortor quis diam luctus, quis tristique tortor ultrices. Integer et lacus a velit efficitur convallis. Morbi enim erat, fermentum vel nulla id, viverra vehicula nisi. Integer non auctor leo, eu convallis massa. Cras eu cursus ligula. Nunc non purus et sem vehicula viverra ut nec nibh.

Quisque posuere purus quis eros auctor efficitur. Etiam mattis vitae nulla et blandit. Nulla a orci magna. Cras ac elit enim. Vestibulum nec nisl metus. Mauris congue velit nec malesuada scelerisque. Sed dignissim, enim vitae semper fermentum, mauris leo vestibulum nisl, in malesuada nibh felis nec dui.

Nullam sit amet tellus eget mi varius commodo. Vestibulum sit amet egestas odio. Nam in ullamcorper quam, nec efficitur augue. Curabitur eget elit in leo eleifend tempor vel lobortis lorem. Duis neque dui, tempus eu sollicitudin ac, lobortis sit amet odio. Morbi eleifend, tellus a varius consequat, enim erat sagittis justo, ac rutrum ipsum augue in leo. Suspendisse non mi ante.

Praesent sed pretium dui, id volutpat tortor. Suspendisse tortor lorem,

vestibulum vitae ullamcorper vitae, tincidunt nec leo. Proin interdum congue blandit. Ut bibendum sagittis leo, nec venenatis urna mollis id. Donec nec erat non justo maximus venenatis. In mollis elit eu odio maximus porta. Vestibulum varius turpis sit amet orci blandit, vitae volutpat erat viverra.

Suspendisse nunc urna, elementum ut purus a, sagittis porta velit. Integer ultricies convallis tortor id pellentesque. Duis et sem a mi bibendum congue. Morbi ut tellus cursus, laoreet ipsum rutrum, condimentum felis. Proin velit mi, ultricies a urna nec, facilisis pretium mi. Pellentesque tristique interdum purus, a facilisis mi tempor quis. Sed finibus venenatis ligula porttitor porttitor. Suspendisse cursus lorem nec velit commodo fringilla. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque a magna quis nunc venenatis vestibulum. Curabitur commodo efficitur ipsum, non ullamcorper tellus. Duis dictum commodo nisi nec venenatis. Donec euismod pulvinar finibus. Suspendisse lorem mi, suscipit quis faucibus ut, luctus in justo. Cras pulvinar arcu ut ullamcorper pulvinar. Aliquam dictum tortor quis diam luctus, quis tristique tortor ultrices. Integer et lacus a velit efficitur convallis. Morbi enim erat, fermentum vel nulla id, viverra vehicula nisi. Integer non auctor leo, eu convallis massa. Cras eu cursus ligula. Nunc non purus et sem vehicula viverra ut nec nibh.

Quisque posuere purus quis eros auctor efficitur. Etiam mattis vitae nulla et blandit. Nulla a orci magna. Cras ac elit enim. Vestibulum nec nisl metus. Mauris congue velit nec malesuada scelerisque. Sed dignissim, enim vitae semper fermentum, mauris leo vestibulum nisl, in malesuada nibh felis nec dui.

Nullam sit amet tellus eget mi varius commodo. Vestibulum sit amet egestas odio. Nam in ullamcorper quam, nec efficitur augue. Curabitur eget elit in leo eleifend tempor vel lobortis lorem. Duis neque dui, tempus eu sollicitudin ac, lobortis sit amet odio. Morbi eleifend, tellus a varius consequat, enim erat sagittis justo, ac rutrum ipsum augue in leo. Suspendisse non mi ante.

Praesent sed pretium dui, id volutpat tortor. Suspendisse tortor lorem,

vestibulum vitae ullamcorper vitae, tincidunt nec leo. Proin interdum congue blandit. Ut bibendum sagittis leo, nec venenatis urna mollis id. Donec nec erat non justo maximus venenatis. In mollis elit eu odio maximus porta. Vestibulum varius turpis sit amet orci blandit, vitae volutpat erat viverra.

Suspendisse nunc urna, elementum ut purus a, sagittis porta velit. Integer ultricies convallis tortor id pellentesque. Duis et sem a mi bibendum congue. Morbi ut tellus cursus, laoreet ipsum rutrum, condimentum felis. Proin velit mi, ultricies a urna nec, facilisis pretium mi. Pellentesque tristique interdum purus, a facilisis mi tempor quis. Sed finibus venenatis ligula porttitor porttitor. Suspendisse cursus lorem nec velit commodo fringilla.

Bibliografia

[1] Latex.

Ringraziamenti

Qui possiamo ringraziare il mondo intero!!!!!!!!!!
Ovviamente solo se uno vuole, non è obbligatorio.