How to fine-tune a Transformer to transcribe historical documents

24/10/24

Sara Ferro, Post Doc @CCHT

Mail: Sara.Ferro@iit.it

@PyDataVenice

Università Ca'Foscari Venezia

iit ISTITUTO ITALIANO DI TECNOLOGIA CENTER FOR CULTURAL HERITAGE TECHNOLOGY

Obiettivi

1. Tipo di applicazione
2. Modello utilizzato
3. Libreria di Hugging Face:
   - Pacchetto “transformers”
   - Moduli utilizzati (e.g., per settare i parametri di training)
   - Codice
4. Le “model cards” e la loro importanza

0. Una breve introduzione del CCHT

Center for Cultural Heritage Technologies (CCHT)

Tre rami:

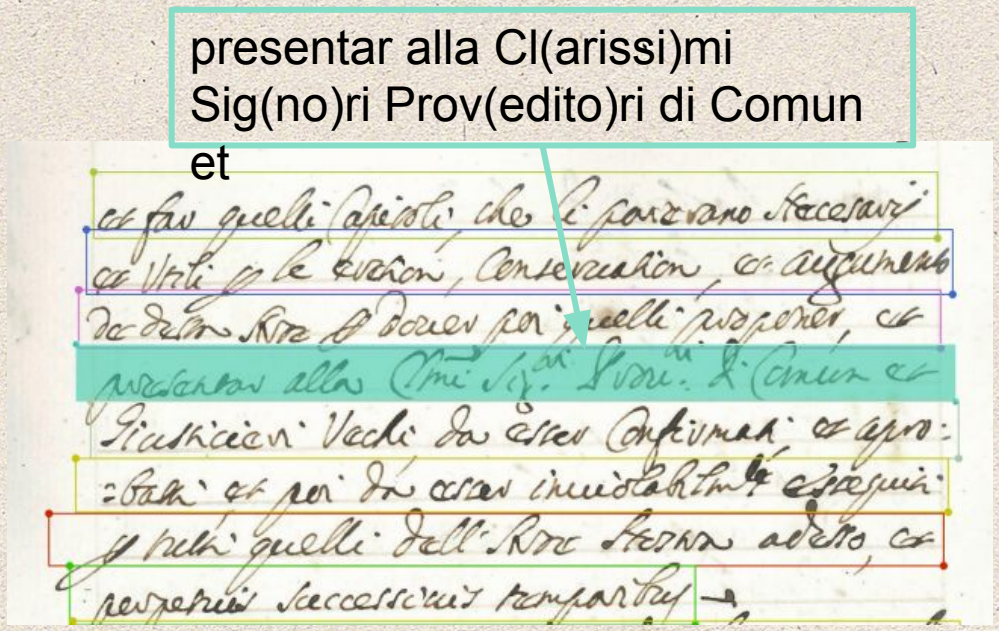
1. *Machine Learning*
2. *Chimica*
3. *Robotica*

applicati ai beni culturali.

RITHMS
Research, Intelligence and Technology for Heritage and Market Security

1. Tipo di applicazione

Optical Characeter Recognition (**OCR**) o meglio **Handwritten Text Recognition** (**HTR**) per la *trascrizione automatica di documenti (storici) manoscritti*



2. Modello utilizzato

- **TrOCR** sviluppato da **Microsoft nel 2021**
- Hanno rilasciato 3 diversi tipi di modelli che differiscono:
  - Tipo di encoder e decoder
  - Pesantezza del modello (# di parametri del modello)

1. *SMALL TrOCR* (62M di parametri)
2. *BASE TrOCR* (334M di parametri)
3. *LARGE TrOCR* (558M di paramtri)

2.1. Pacchetto “transformers”

```
import transformers
```

Transformers (https://huggingface.co/docs/transformers/en/index)

Integra i modelli migliori che sono stati rilasciati dalla comunità scientifica basati sull’architettura del Transformer

Da APIs e strumenti per scaricare facilmente modelli (basati su Transformer) che sono stati pre-addestrati

L’utilizzo di modelli pre-addestrati:

1. Riduce il costo computazionale per ottenere un determinato obiettivo
2. Riduce la “carbon footprint”
3. Riduce il tempo necessario per addestrare il modello, non dovendo partire da zero

2.1. Pacchetto “transformers”

1. **Natural Language Processing** (**NLP**) – e.g., named entity recognition, generazione di testo e modelli di linguaggio
2. **Computer Vision** (**CV**) – e.g., classificazione di immagini e segmentazione di immagini
3. **Audio** – e.g., riconoscimento automatico della voce e classificazione dell’audio
4. **Multimodale** – e.g., OCR e estrazione dell’informazione da documenti scansionati

2.1. Modulo principale

```
from transformers import VisionEncoderDecoderModel, Seq2SeqTrainer,
Seq2SeqTrainingArguments, TrOCRProcessor
```

VisionEncoderDecoderModel

Classe utilizzata per inizializzare un modello da immagine a sequenza con:

1. Modello di visione pre-addestrato
2. Modello di linguaggio di tipo autoregressivo

class transformers.VisionEncoderDecoderModel

```
( config: Optional = None, encoder: Optional = None, decoder: Optional = None )
```

2.2. Moduli utilizzati del modulo “transformers”

- Sia l’encoder sia il decoder possono essere inizializzati “a piacere” scegliendo un encoder e decoder tra i disponibili che sono in HuggingFace
- `from_pretrained()` è la funzione per caricare i modelli pre-addestrati
- Tale classe eredita dalla classe `torch.nn.Module`

si può quindi facilmente passare a PyTorch e viceversa

2.2. Moduli utilizzati del modulo “transformers”

- `Seq2SeqTrainer` è una classe che serve per fare il training di modelli in *ingresso una sequenza e danno in uscita una sequenza*
- `Seq2SeqTrainingArguments` è la classe che serve per *passare gli argomenti al trainer*
- `TrOCRProcessor` non è un “modulo comunemente usato”. E’ stato implementato da Microsoft per creare un unico modulo che avesse:
  1. *Visual image processor*
  2. *Tokenizer*

  in un *unico oggetto*

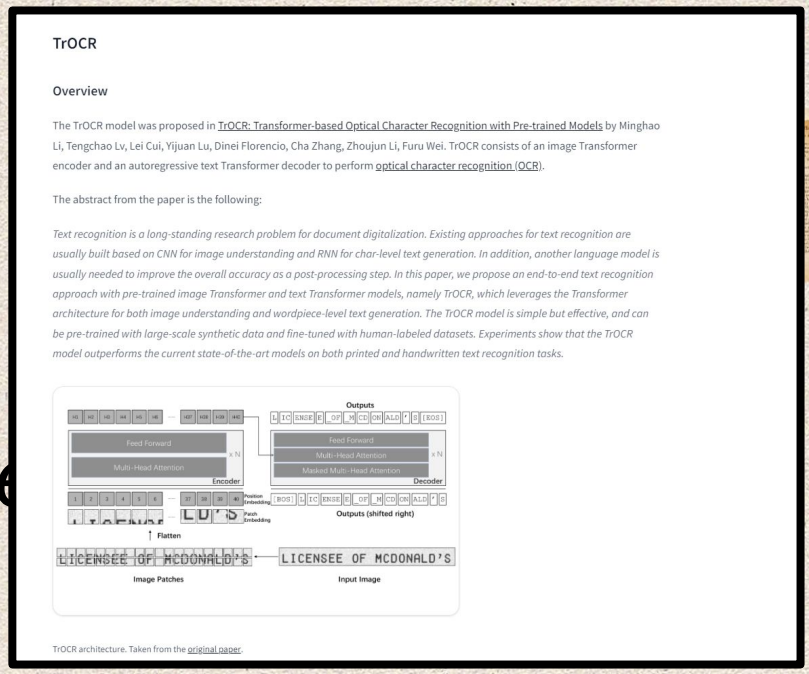
Codice

Apriamo Google Colab e vediamo sia l'inferenza sia il training nella pratica!

https://colab.research.google.com/

Le “model cards” e la loro importanza

Molti modelli hanno informazioni nelle cosiddette “*model cards*” che sono scritte dagli sviluppatori dei modelli per *documentare come poter utilizzare il modello*



TrOCR

Overview

The TrOCR model was proposed in TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models by Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, Furu Wei. TrOCR consists of an image Transformer encoder and an autoregressive text Transformer decoder to perform optical character recognition (OCR).

The abstract from the paper is the following:

Text recognition is a long-standing research problem for document digitalization. Existing approaches for text recognition are usually built based on CNN for image understanding and RNN for char-level text generation. In addition, another language model is usually needed to improve the overall accuracy as a post-processing step. In this paper, we propose an end-to-end text recognition approach with pre-trained image Transformer and text Transformer models, namely TrOCR, which leverages the Transformer architecture for both image understanding and wordpiece-level text generation. The TrOCR model is simple but effective, and can be pre-trained with large-scale synthetic data and fine-tuned with human-labeled datasets. Experiments show that the TrOCR model outperforms the current state-of-the-art models on both printed and handwritten text recognition tasks.

TrOCR architecture. Taken from the original paper.

Conclusioni

- *Huggingface ha come obiettivo il rendere chiunque capace di utilizzare e affinare modelli ottimi* (state-of-the-art models)
- E' *facile da usare*
- Esiste il *forum per fare domande*, inoltre esiste un *canale Discord*!
- *Leggete la "model card" prima di utilizzare un modello!... Qualche buona anima ha fatto la fatica di scriverla!*

Thanks!

Abbiamo posizioni aperte:

Research Fellow in Digital Twins

In più…

Sei uno student che deve fare la ***tesi magistrale****?*

Se vuoi puoi farla collaborando con noi su progetti che utilizzano:

1. ***NLP***
2. ***Clustering***

Sentiamoci!
Sara.Ferro@iit.it