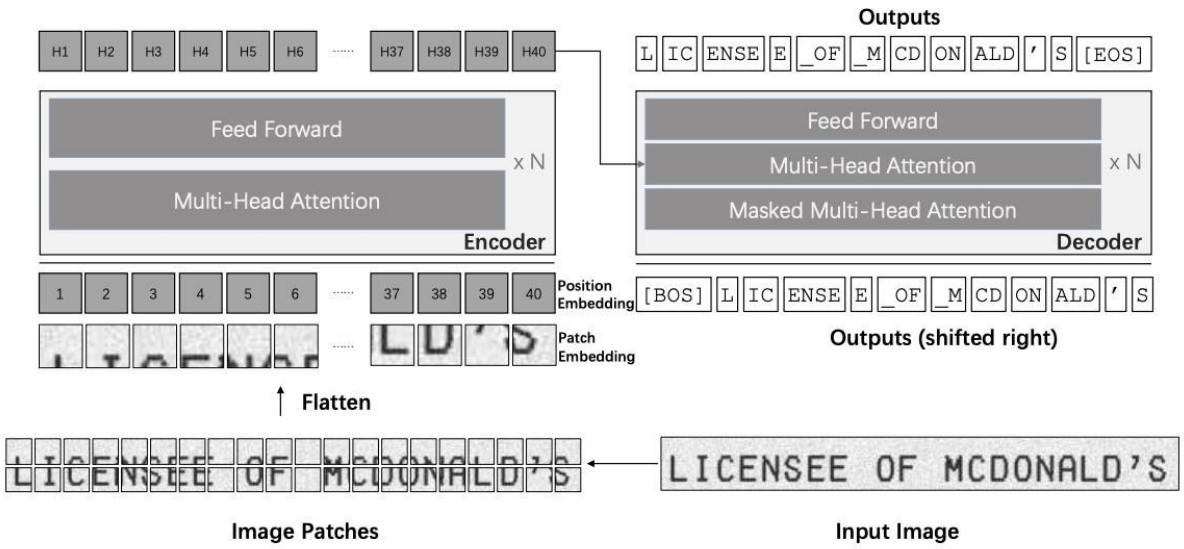
Optical character recognition For Arabic Text

This is a fine tuning repo for TrOCR on Images that contains arabic text This file is

organized as following:

- `TrOCR Architecture`
- `Dataset`
- `pipeline`
    a. Initial Dataset
    b. Data Cleaning
    c. Preprocessing
    d. `Training Dataset`
    e. Training
    f. Evaluation

The Model Architecture



- TrOCR paper: https://arxiv.org/abs/2109.10282
- TrOCR documentation: https://huggingface.co/transformers/master/model_doc/trocr.html
- TrOCR Tutorial : https://github.com/NielsRogge/Transformers-Tutorials/tree/master/TrOCR

Pipeline

Initial Dataset

- data contains Faulty Images with croped characters that needs to be removed
- All Images contains one line and have the same background (distribution)
- sentence length is falls between `9-13 tokens`

Data Cleaning

The dataset had croped text that needed to be removed from the training data this was done using

1. Detect the Height of the characters
2. Compare Maximum character Height to choose `threshold = 17` 3. Remove the

Faulty Images `8%` leaving `92%` for training

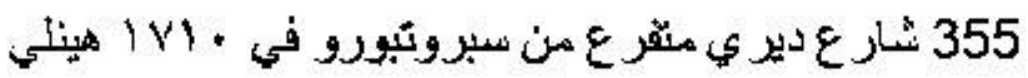
example:

Data Preprocessing

Removing Background and enhance the characters

1. convert image to greyscale
2. Binarize image `threshold = 110`

Output:



Training Dataset

Due to Limited Time and Computational Power espicially RAM A random Sample was taken from the initial data with :

- training 1400 sample
- evaluation 600 sample

Training

The training used HuggingFace's Seq2SeqTrainer :
https://huggingface.co/docs/transformers/main_classes/trainer#transformers.Seq2SeqTrainer

choosing the encoder and decoder for the Task

```
# choosing feature extractor and tokenizer
feature_extractor =
AutoFeatureExtractor.from_pretrained("google/vitbase-patch16-384")
decoder_tokenizer = AutoTokenizer.from_pretrained('xlm-roberta-base')
processor =TrOCRProcessor(feature_extractor=feature_extractor,
tokenizer=decoder_tokenizer)
```

Train parameters

Note : some of these values were chosen to reduce Memory and computational power due to hardware limitation

- `maxlength for seq = 20`
- `batch size =3`
- `epoch =3`

Evaluation

Character Error Rate (CER) metric for evaluating the performance of a sequence-to-sequence model
https://huggingface.co/spaces/evaluate-metric/cer