Fine-Tuning mBART for Arabic to English Machine Translation using LoRA (low rank adaption)

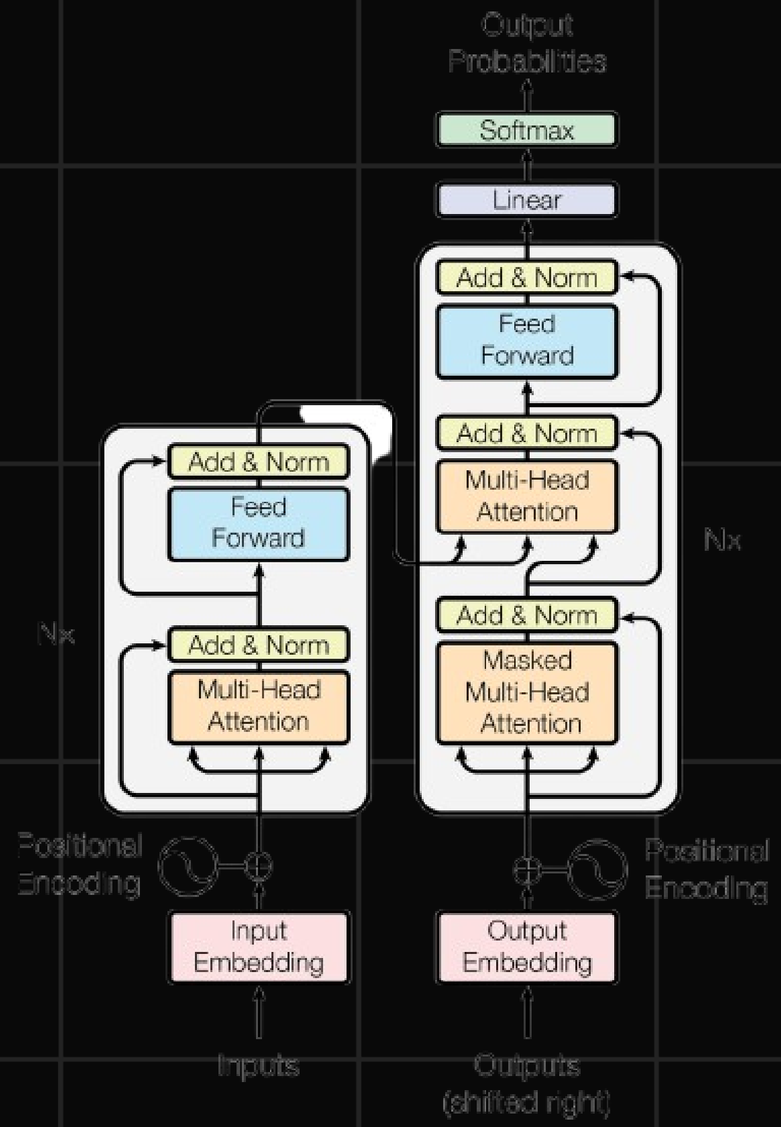
Translation

AR → En

ElZozat +2

USING M-Bart pre-trained Model from Huggeing face

It’s an LLM with 611M parameter to handle 50
Different Language Translation
Based On the transformer Architecture

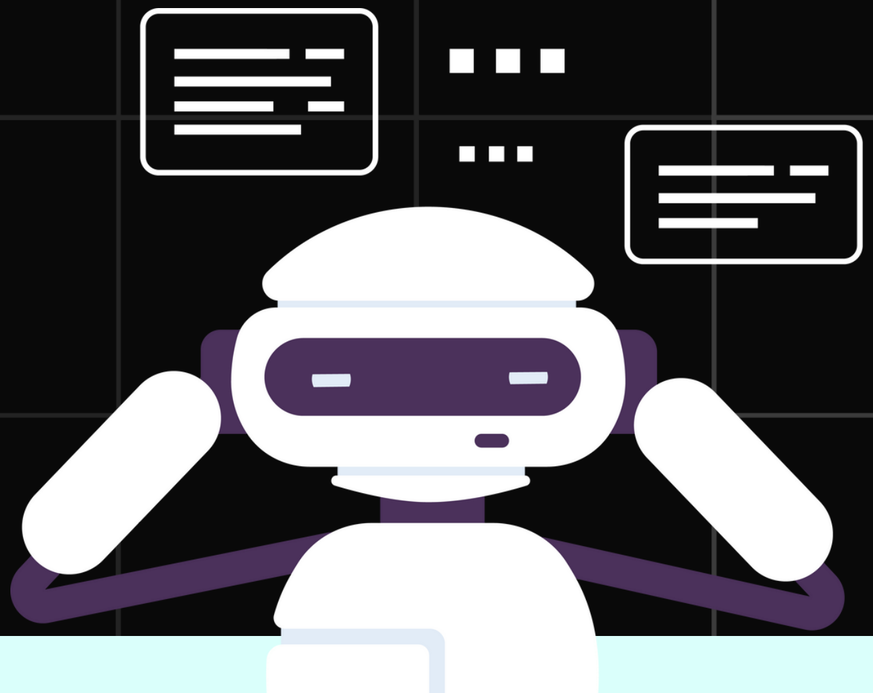


ElZozat +2

PEFT

(Parameter -Efficient Fine -Tuning)

Since We have got a huge number of parameters for the base model to fine -tune so We need extreme computing resources to Fine tune the model so we used PEFT (LoRA – Low Rank Adaptation) to overcome this



ElZozat +2

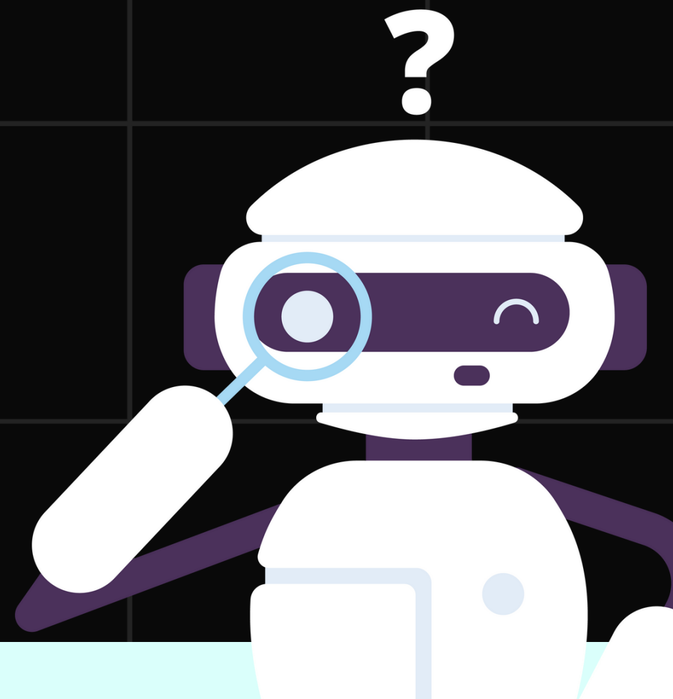
Link:
https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

DataSet

Link:
https://huggingface.co/datasets/ymoslem/CoVoST2-EN-AR-Text

269,380 Row (AR -EN)
Cleaned and normalized text
269,380 sentence pairs (train: 240k, validation: 15k, test: 14k)

We Used only 150K Sample from the data
Splitted into train (120k), validation (15k), test (15k)



ElZozat +2

Model Parameter

M-Bart Base Model

- 611M parameter
- 12 (24 total)encoder/decoder layers
- 1024 hidden dimension
- 16 attention heads

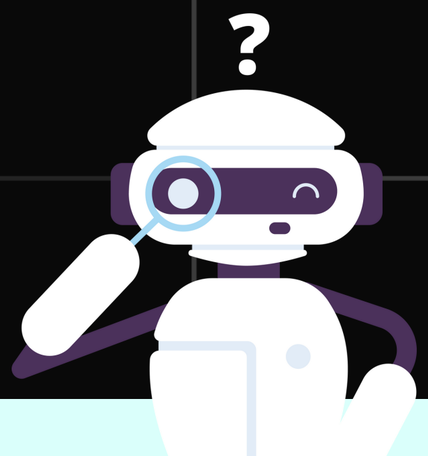
LoRA Adaptor

Applied to attention layers (Query, Key, Value)

- Rank = 16
- Alpha = 16
- Dropout = 0.05
- Number of trainable param : 3.5M

Training specs

- Epoch: 5
- Batch size: 8
- Learning rate: 3e -5
- Seq length: 128 tokens



ElZozat +2

Training Process Steps

Loading the Dataset from HF& preprocessing

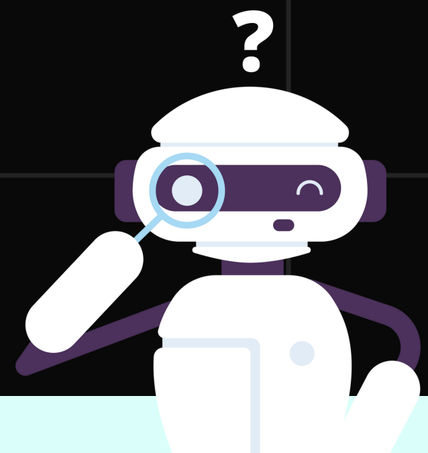
Configure The tokenizer

Configure The LoRa Adaptor

Configuring the trainer args and then start training the model

Evaluating The Model

Saving the Model To HF



ElZozat +2

Evaluation
for (15K Testing Sample)

BLEU - SacreBLEU - chrF - chrF++

Base Model

```
Example Translations:

Source (Arabic): ترعى جمعيه Skeptics سلسله من المحاضرات في معهد كاليفورنيا للتكنولوجيا.
Reference (English): the skeptics society sponsors a lecture series at the california institute of technolog
Prediction (English): His group, Skeptics, holds a series of lectures at the California Institute of Technol

Source (Arabic): تم استرداد ملاعق الفتره الرومانيه من الحفريات في مدينه لندن.
Reference (English): roman period spoons have been recovered from excavations in the city of london.
Prediction (English): The Roman bath lamps were recovered from the excavations in London City.

Source (Arabic): ورغم ذلك يمكن تسويقه وبيعه في الولايات المتحده.
Reference (English): nevertheless, it can be marketed and sold in the united states.
Prediction (English): And yet it can be sold and traded in the United States.

Source (Arabic): اشار الى ان السبب هو ضغوط السفر المبالغ فيه.
Reference (English): he cited the pressure of too much travel as his reason.
Prediction (English): He pointed out that it's because of the overwhelming travel pressures.

Source (Arabic): تدعم المجموعه فرض ضرائب متزايده على المتسببين بالتلوث ومنح حوافز لاصحاب الممارسات المفيده بيئيا.
Reference (English): the group supports increased taxes for polluters and incentives for environmentally ber
Prediction (English): The group supports increasing taxes on polluters and incentives for environmentally-fr
Quick BLEU Score (15K samples): 0.16
Quick SacreBLEU Score (15K samples): 16.00
Quick chrF Score (15K samples): 45.38
Quick chrF++ Score (15K samples): 42.74
```

Fine tuned model

```
Example Translations:

Source (Arabic): ترعى جمعيه Skeptics سلسله من المحاضرات في معهد كاليفورنيا للتكنولوجيا.
Reference (English): the skeptics society sponsors a lecture series at the california institute of te
Prediction (English): skeptics sponsors a series of lectures at the california institute of technolog

Source (Arabic): تم استرداد ملاعق الفتره الرومانيه من الحفريات في مدينه لندن.
Reference (English): roman period spoons have been recovered from excavations in the city of london.
Prediction (English): the remains of the roman period were recovered from the excavations in london.

Source (Arabic): ورغم ذلك يمكن تسويقه وبيعه في الولايات المتحده.
Reference (English): nevertheless, it can be marketed and sold in the united states.
Prediction (English): however, it can be traded and sold in the united states.

Source (Arabic): اشار الى ان السبب هو ضغوط السفر المبالغ فيه.
Reference (English): he cited the pressure of too much travel as his reason.
Prediction (English): he pointed out that the reason was the excessive travel pressures.

Source (Arabic): تدعم المجموعه فرض ضرائب متزايده على المتسببين بالتلوث ومنح حوافز لاصحاب الممارسات المفيده بيئيا.
Reference (English): the group supports increased taxes for polluters and incentives for environmenta
Prediction (English): the group supports increased taxes on polluters and incentives for environmenta
Quick BLEU Score (15K samples): 0.34
Quick SacreBLEU Score (15K samples): 34.48
Quick chrF Score (15K samples): 59.16
Quick chrF++ Score (15K samples): 57.89
```

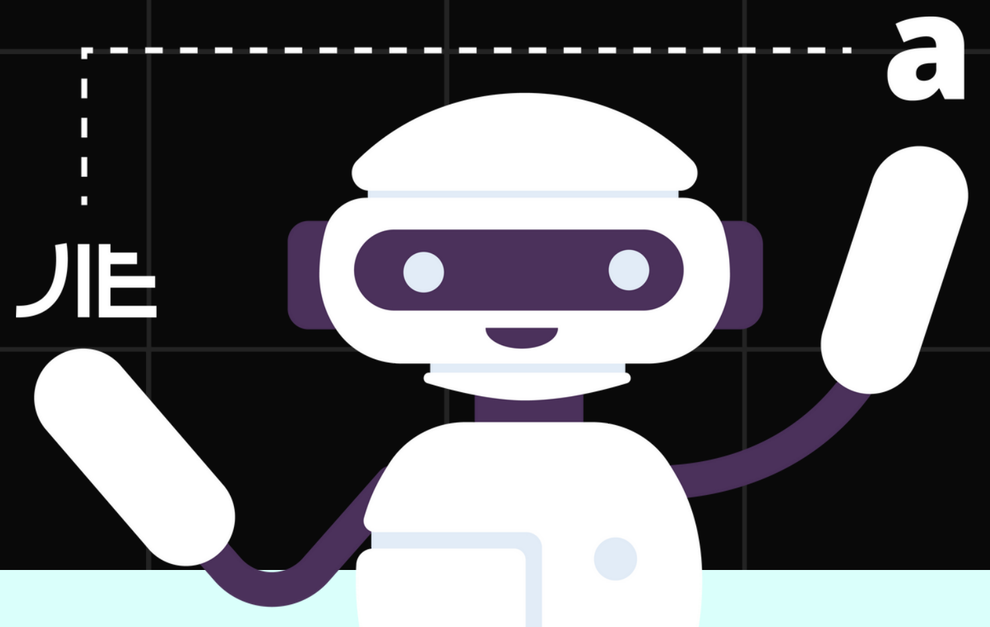
ElZozat +2

Model Limitations

- Rare word translation issues
- Longer sentence quality degradation
- Not Resources Friendly (needs a High GPU)

Arabic -specific challenges:

- Dialectal variations
- Complex morphology



ElZozat +2

Our Team

Ziad Adel Farghaly

Ziad Waleed Mohamed

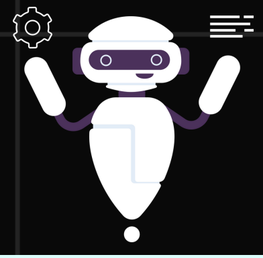
Ziad Tarek Galal

Ziad Atef Ali

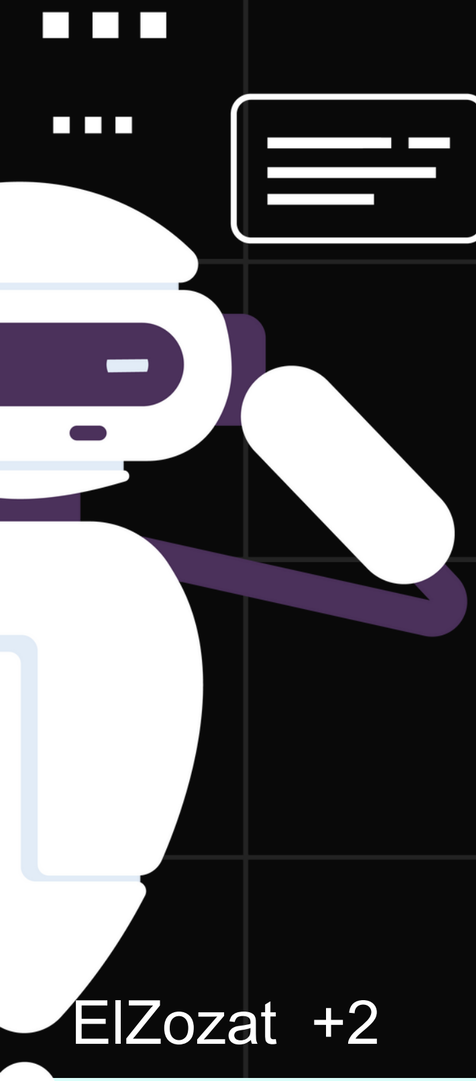
Ziad Ahmed Ali

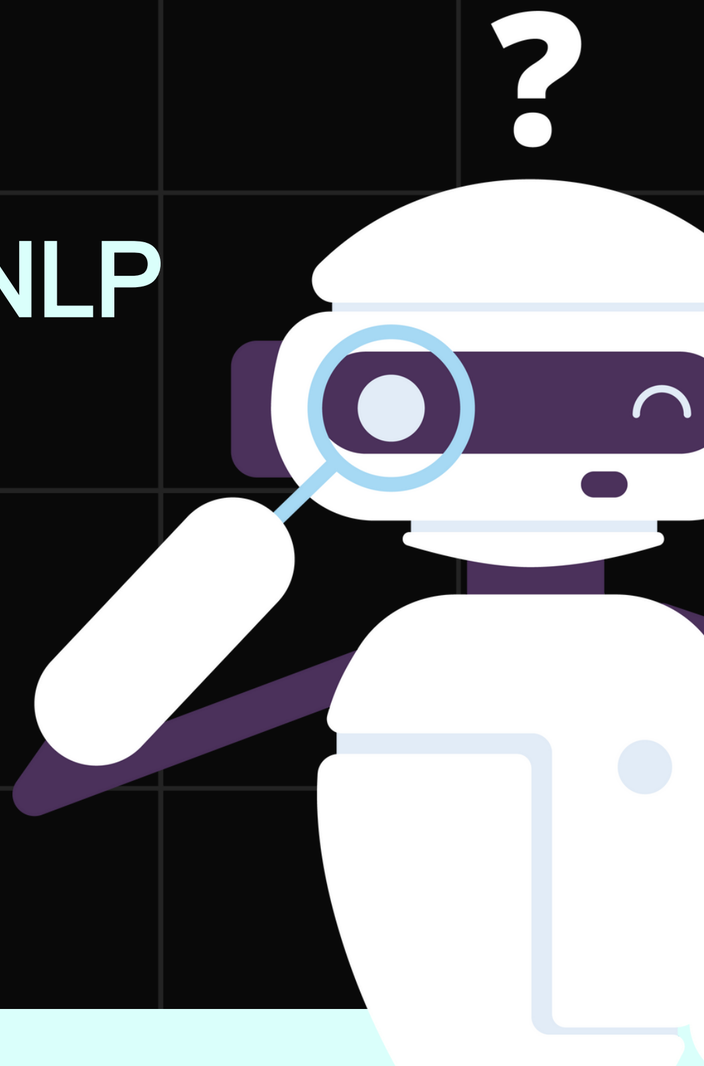
Mostafa Gamal Eldin

Basel Waleed Hamed



From Elzozat +2 To all The NLP Developers





ElZozat +2

Thankyou

https://github.com/ZiadWaleed2003/LoRA-Fine-Tuning-mBart-for-Machine-Translation

ElZozat +2