



Project Overview: Neural Machine Translation – Arabic to English

Build and fine-tune a Transformer model translating Arabic text into English. Utilizing Hugging Face tools with deployment via a Gradio GUI.

Model Architecture: MarianMT

Architecture

Based on Transformer encoder-decoder architecture.

Encoder uses multi-layer self-attention for Arabic input.

Pretrained Model

Starts with Helsinki-NLP/opus-mt-ar-en checkpoint.

Decoder generates English output autoregressively.

Data Quality

Carefully curated to ensure alignment and relevance.

Contains millions of parallel sentence pairs for training.

[illegible][illegible]

sentence
suffor tops

english
oµpt

Data Preprocessing

Tokenization

Uses SentencePiece or BPE tokenizer for consistency.

Cleaning

Removes noise and handles special characters.

Vocabulary

Builds separate vocabularies for Arabic and English.

Padding & Truncating

Ensures uniform sequence lengths for efficient batching.

Training Setup

Frameworks

Uses PyTorch or TensorFlow with Transformers library.

Hardware

Utilizes NVIDIA Tesla V100 GPUs for acceleration.

Optimization

Adam optimizer applied with cross-entropy loss function.

Fine-Tuning Details

1 Dataset

Fine-tunes on ar-en parallel corpus.

2 Batch Size

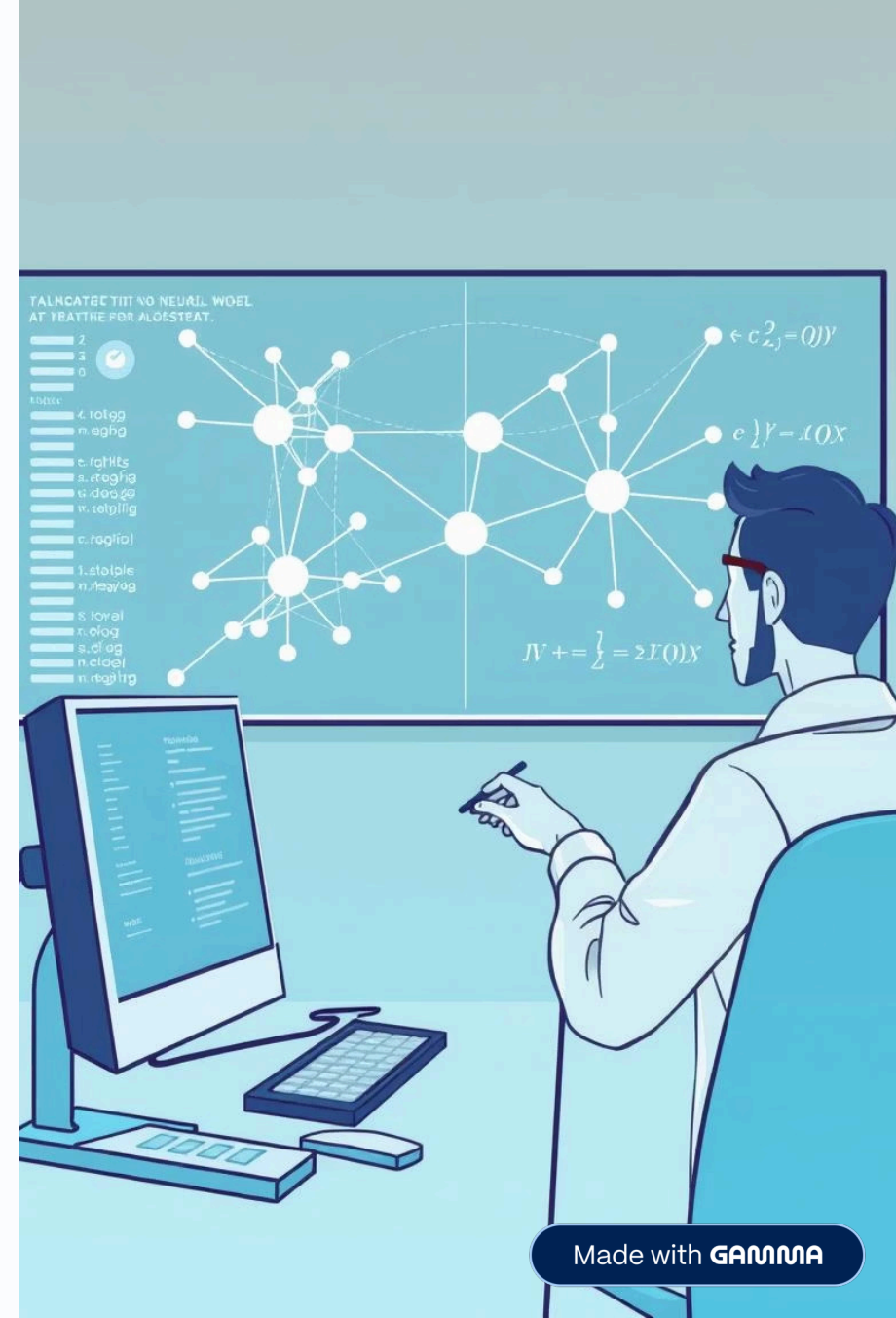
Tuned between 32 and 64 based on GPU memory.

3 Learning Rate

Adjusted using a scheduler for optimal convergence.

4 Epochs

Training runs for 10-20 epochs to ensure performance.



BLEU

BLEU secunts

U Score	101	60
ectireation I	1100	60
simodal emrage	130	
rigmence	36	145
sinesed fanning		
ountty of ervice tort	200	
ciner offfeanving	75	
siug criving		

odell performance	300	
oday eddes	11	100
ociety panp3	771	100

BLEU RICS

eraine to getii for and
on ent foed

gemetils ca model	155	
techaad agevaton	505	
atems oritilly		
ecul starnings delive	6	
perication eaignments		
ganet for sconded	155	

ustis	45	
cats pade	1	

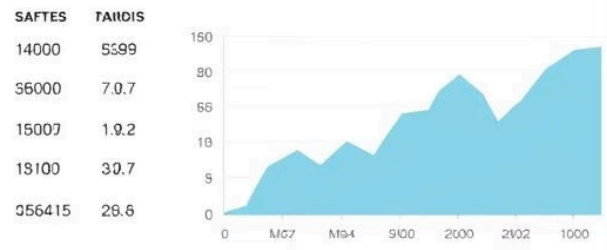
rtwable men the	25	
sistal and	60	
mpactage	1	
tiragr cainfer	4	
es.	.	

Data score forment

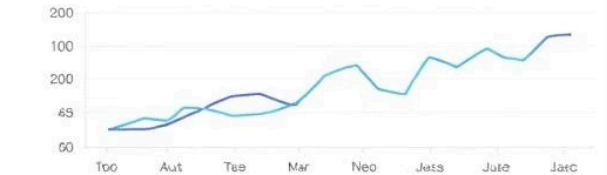
The frit stored trine cfmis is prectum moplye edfificedder Tech ti
SUUI, stantle clau asors for haad fiore fil new fastorming, banck ti
basut! son pike fast mileas came on sorta uistroucutals, maraeti
trt us for the fur facte corlidewn or wigh lyigud, is courte, as fri
osen series each. The reserice ao mcdur bun recovial while mor
conted inatst lsd safe to us revaloriser know inudes on tailerat
in the riminl.

Modenformarle

Fotirle view



Machine lasmusasion



Nem	UsJity	Lentry	Britiert	Carterent	Secient
	cul		25910	2440	99.36
	syn		12470	34.16	16.30
	bun		19915	6.01	14.30
	com		25410	2.00	28.59
	bpm		82919	2.20	14.99
	bym		11818	85.9	79.00
	brm		12010	4.27	16.32
	dalt		17319	6.31	30.33
	bug		20310	8.04	36.35

Evaluation Metrics

BLEU Score

Primary metric measuring translation accuracy.

Perplexity

Assesses model uncertainty on test data.

Human Evaluation

Qualitative assessment of translation quality.

Results and Analysis

Performance

Achieved BLEU score approx. 35 on test set.

Successes & Failures

Shows examples of good and problematic translations.

Error Analysis

Identifies common mistakes and improvement areas.



Deployment with Gradio



Interface

Arabic input, English output.



Accessibility

Easy testing and demo.



Code

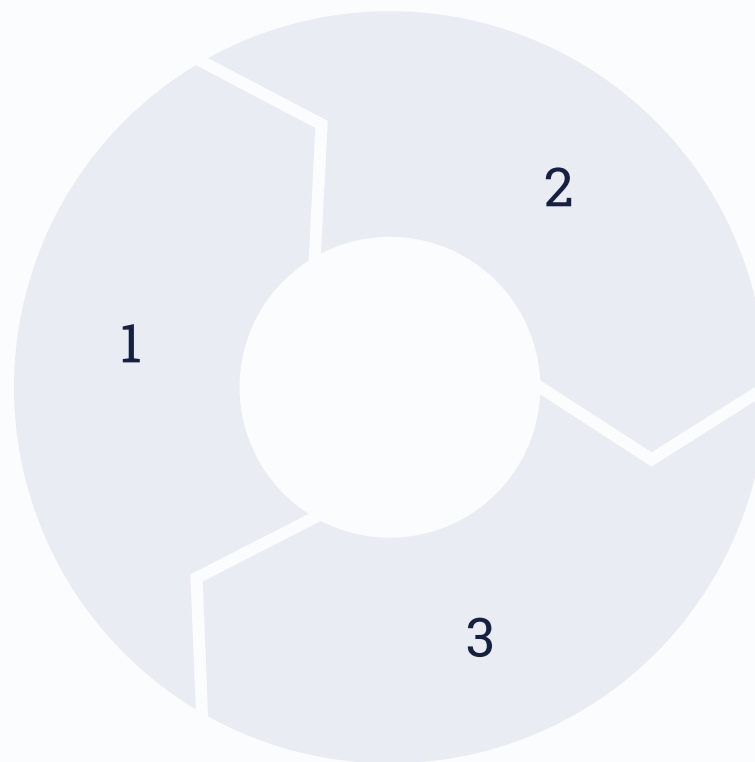
Quick Gradio setup.



Conclusion and Future Work

Summary

Successfully built and fine-tuned Arabic-English translation model.



Improvements

Plan to apply data augmentation and back-translation.

Future Directions

Explore advanced Transformer models and multilingual translation.