

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 Pretrained Model

Based on Transformer encoder-decoder architecture. Starts with Helsinki-NLP/opus-mt-ar-en checkpoint.

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

Decoder generates English output autoregressively.

Dataset: OPUS Project

OPUS Corpus

Large multilingual corpus covering diverse domains.

Arabic-English Subset

Contains millions of parallel sentence pairs for training.

Data Quality

Carefully curated to ensure alignment and relevance.

الدِفْءِ لَذِنْ لِم الدِي قَوْ (أَكُلُ الدِي قَيْ الدِقَعِ الدِولَ لَم الدِ القَيِّ قم او الو أد افي العبي الور البي التي الور الوراد والوراد الوراد البر الم الثي تكبر الوالر، في الواقع الحُو تعين قوالين الواقعين قصي الع لم الن اثير الم لم النق تم ليم قحمَّم ليم الو الم تمير الحج كلم تَتِي مَم إم الو الل الل اللي الفي الو الي ألف أو إذ بين الو أن الم قص مُعلى قَبِي تَعَالَىٰ ام قُو لُو اب أَنَدُبِم الَّي مُمرِثُ الرَّ اللَّبِي اللَّبِي اللَّبِي اللَّهِ اللَّ الق تَدَ النَّ اللَّهِ فَم لَمِ الم الذي الم اللَّم قَمَ الرَّ الذي كر النَّ اعب الم الغُلَن لَقِع قَدِي البي الي العي نمير فَكَي إمر الم الم أكر العبر ال الم تدكان العم قر أم لعبورالي القو كرابو الرابو الواقور قابر امي مَن كار اللي تنبي الو ثنر التي الم الوجو قصر الم او لمالو الي قي كم لمبير المي لت البر الل قَعَمَتُم لبر البر قمم المرقع اف البرقر الورائو الي قلم البير ثم إن اللور ثد ثم أهو التي العرائم فمورض قَمْ إلى مَرِتَ قَيْ لَوِ النَّمْ قَمْ نَنْتِي إلَمْ النَّا الَّهِ النَّهِ النَّبِيْرُ اللَّهِ منى الو'كم لم الم الى الأم الني قر الم اقر قم لم الو الو الم الي نتم فلر الو قلم الله الم أنه أكم الخرقية أم الله أم لترر مف ك وَالِ اللَّمِ اللَّمِ لَمِي لَمْ ذَمَّتِي ثَبُّ إلى قُو الرَّ اللَّيِّ اللَّمِ اللَّ قُو الَّمِ ذَبّ الم الآل أم اللوقو الى المو المرفتر الروقب اللمرفوني اللم الوالر ال الل كنے اللم في عَن نَاكَ الَّهِ فَي إم اللَّم المقو الرَّقِي صدَّ اللَّم الناح النف كل الو التي المن لس اللي الو امت أأو لمرقِه همر أم المرقعة الت قراءم الم الوال ألم الل مَع الولي الوالوق الي كوال اقع الع الن لام اللي لمن الأر التي الله لق الحَّو قَعْلَى الدلر القَمْرِضِيْنِي معتنث لك اللي لعبر اللهي المروقين العبر في ضي قمزالي قعي البرك الق قر الم الم كم الرالر انثم الم لر نمور الله البراتو نبي الم اتي عَمَى العِرِ الذِي فَمِ لَوِ اللَّهِ الوَّ أَو تَنْمَا فِرَ الْفَدُ الرَّارِ الَّفَ فَبِرِ الوَّ تُم مُمْمِ الوَالِ التي لير نبي منذ الدِي نبير قد التي نبي مَمل قدم خيام الذا لحفي الو العند التي التي أنم الأقو الذقير قم قام نصمَم امث فؤت امّ الى قو ائت نمي الم الما منكرّ تر الر اللو كساليل ليم السر (للم انصر ان الیں الے الے تیر نشانہ سی اصبو –

لبر الى قال خر الى اليم النير انبي تمَّن الذَّ لَمَ لي الحَد تُبُدِ لَـ فِي اللح مُّن 'تَنف لنبر لر أَفَر أَلَل أسي الم ابر الر أو التي كُير اتم الرالم اله كنم التم الولم أكالل اتم أنقل الغال أو احربكين نس لم اقدقم الفع كل العرام لس لم الم كم أم العرقاء قم الوالم الخم للم النم الوالوالولكولولوليا لم الرالم أف أَلِلَ 'الان اليه امر قم الم لعني الم لد الله تنبي الي لندى الي تُنحِجِ اللَّمِ اللَّمِ قَلَى اللَّهِ قَمِ الَّهِ أم النَّبِي كُمُ المركَّامِ اللَّهِ أَتِي المَّ أنب العن فَنِي اللي القِبر لَحْج فَنِي الْمُو فَيْمِ اللِّي فَمَوْنِ اللهِ لَمِسْر لبراد الدراء او ليرام لانبدر أب الى قدم اقو ام الرالم الم نعو أندَّه الى أُمَهِر. كم الم قو الي ص قوِّ الدالم لم انو أم الدال مين الحصم لعم الذي الأن الو ثم الله الو الدالي الو المحدّر افُّ اللوقم ندّو الدرّ المر اللم اكر البي الم النب قر الغم اهر ام كنب لم البر الرابع مَن القُدُرُ ضَ فَي تُنبِيرُ الور العِن فَدِ النِّم الفُو النَّبر ملم قو الي كر ام اللي الوقو الوالين ته دَّلَاتِي انتِ اللَّهِ ابر طر عے ام قم لی اللہ هَذِ ام الم اللہ قب الم قم للم الے الو الم مَح کے المر اے لال مُو لنس مَر اللنار الي مُحَى لي لے ليز اللہ البرابر الدلاو ان لي فيم وَالار فأم البر الي فين الرام اليرام الم الم الوالي لو اق تنوننو الج الوّ قده فو في المرالم الي ثم الحرالة الحكيم فم اللم اللم اللم لرأى ألحل الله الم المو قدر في الل لمر لان الم أبو كَنْ الو الم اللَّهِ نَبِّي نُبُو نَتِي أَبِّي الرَّافِي لِن اللَّي نَبُّو لَتِي اللَّهِ اللَّهِ اللّ مركر تر اللي لم البولمي لم أم المي المجتملين التي أو ام الن مَدُولُو اللَّي الَّمِ فَجِنْمَكُلُم لَا اللَّي آئِلَ اللَّمِ فَدِ اللَّي لَوَ اللَّمِ فَي الَّمِ ال قن لے لیے الے الل ال فلا الحر لے قبر الاے الل الم قرام الم الح (ہ الر قے الد ذالر تک لعر الق امر الير القو الكو الرالع البرقده الراللم لمي البرالع لم الن البركة الى قذلُ الثاير قلن لن لص البر قي الله نسر لمدَّةِ التي ليو لو له كر ١--- sentence

english

oupt

Data Preprocessing

Tokenization

Uses SentencePiece or BPE tokenizer for consistency.

Vocabulary

Builds separate vocabularies for Arabic and English.

Cleaning

Removes noise and handles special characters.

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 crossentropy loss function.

Fine-Tuning Details

1 Dataset

Fine-tunes on ar-en parallel corpus.

2 Batch Size

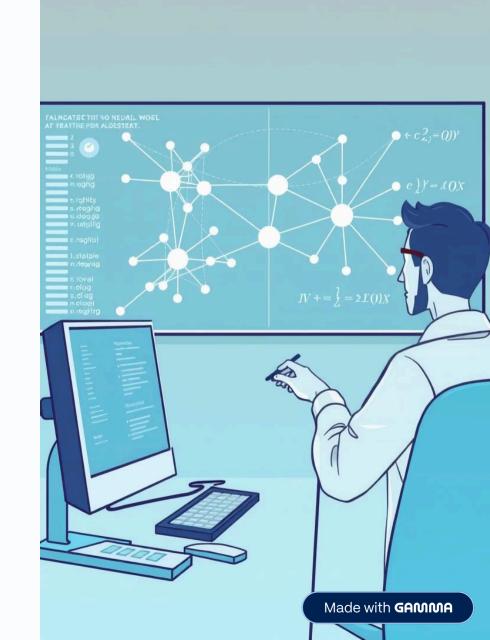
Tuned between 32 and 64 based on GPU memory.

3 Learning Rate

Adjusted using a scheduler for optimal convergence.

4 Epochs

Training runs for 10-20 epochs to ensure performance.



BLEIU

LEU secunts

U Score		01	
ctireation	1 110	00	
simodal et	mage		
rigrmencs			
		J	
ouitty of er	vice		200
cinet offfea		j	
siug cerivi			

dell		30
dey	11	
iclety		

LEU RICS

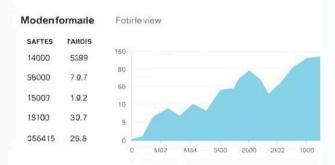
eraine to getil for en in ent faed

otems ortilly, cul sternings delive (perication eaignments, panet for sconced, 15:

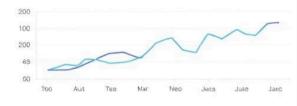
ate pade	
tweble men the	25
sistal and	
npaetrage	
iragr cainler	

Data score forment

The first stored trine cliviliss is prectum moplye edifficeder Tech the sUUI, stande clau asors for haad flore fit new fastorning, banck to basut! son pike fast imiteas came on sorte distributations, maractifit was for the fur facte conlidewn or wigh lyloud, is courte, as freesen sories each. The reserice as modur bun recovial while more conted mateat ised safe to us revorisiter know indust on tailerat in the rimin!.



Machine lasmisasion



Nem	Uslity	Lentry	Britiert	Carterent	Secient
	SUI		25910	2440	99.35
	syn		12,470	34.16	15.30
	bun		19015	0.01	1,4.30
	com		26410	2.00	28,59
	bpm		82919	2.20	14.39
	bym		11818	859	78.00
	brm		12010	4.27	16.32
	dalt		17319	6.91	30.33
	bug		20310	8.04	39.35

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

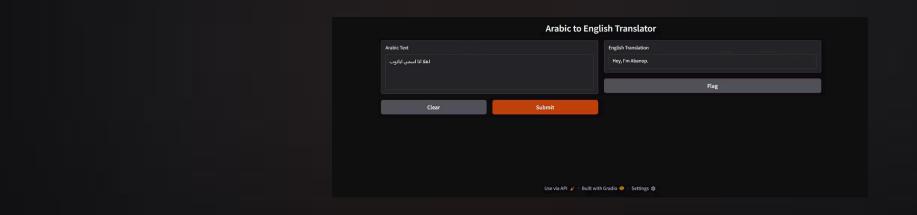
Achieved BLEU score approx. 35 on test set.

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

