LANGUAGE TRANSLATION - Transformer Based Attention Model

Submitted by

Ayushi Kumari-13000118110
Kumar Saurabh-13000118094
Abhijeet karmakar-13000118141
Abhay Kumar Choubey-13000118142

Submitted for the partial fulfillment for the degree of Bachelor of Technology in Computer Science and Engineering

Techno Main Salt Lake,
EM 4/1, SaltLake, Sector V, Kolkata – 700 091

CERTIFICATE

This is to certify that the project entitled “Language Translation-Transformer based Attention Model” prepared by(Ayushi Kumari-13000118110 ,Kumar Saurabh-13000118094,Abhijeet karmakar-13000118141 ,Abhay Kumar Choubey-13000118142).,) of B.Tech (Computer Science & Engineering), Final Year, has been done according to the regulations of the Degree of Bachelor of Technology in Computer Science & Engineering. The candidates have fulfilled the requirements for the submission of the project report.

It is to be understood that the undersigned does not necessarily endorse any statement made, opinion expressed or conclusion drawn thereof, but approves the report only for the purpose for which it has been submitted.

-------------------------------- --------------------------------

(Signature of the Internal Guide) (Signature of the HOD)

------------------------------ ------------------------------

(Signature of External Guide, if applicable) (Signature of the External Examiner)

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor of the department of Computer Science and Engineering, whose role as project guide was invaluable for the project. We are extremely thankful for the keen interest he took in advising us, for the books and reference materials provided for the moral support extended to us.

Last but not the least we convey our gratitude to all the teachers for providing us the technical skill that will always remain as our asset and to all non-teaching staff for the gracious hospitality they offered us.

Place: Techno Main Salt Lake

Date:

………………………………

………………………………

………………………………

………………………………

Contents

1 Introduction 5
Abstract 5
Problem Domain 5
Related Studies 5
Glossary 5
2 Problem Definition 7
Scope 7
Exclusions 7
Assumptions 7
3 Project Planning 8
Software Life Cycle Model 8
Scheduling 8
Cost Analysis 10
4 Requirement Analysis 11
Requirement Matrix 11
Requirement Elaboration 12
5 Design 12
Technical Environment 12
Detailed Design 13
6 Implementation 17
Implementation Details 17
System Installation Steps 18
System Usage Instructions 19
7 Test Results and Analysis 20
8 Conclusion 23
Project Benefits 23
Future Scope for improvements 24
9 References / Bibliography 25

1 Introduction

Abstract

The scientific community is rapidly generating various algorithms on Machine Learning that have helped in various field of image processing, translation and many more. One of the most dominantwork in the field of data science was proposed in 2017 which produced remarkable result. The research paper tells us about the implementation of transformer based attention model. Attention is a concept that helped improve the performance of neural machine translation applications. We are implementing this model to show the translation of language which can further be implemented for the protein sequence analysis.

Problem Domain

In this project, we are translating from Germany to English. Say the following sentence is an input sentence we want to translate:**”The animal didn't cross the street because it was too tired”.** What does “it” in this sentence refer to? Is it referring to the street or to the animal? It’s a simple question to a human, but not as simple to an algorithm. In this project, we will identify this .

Related Studies

Through our study we came to know that ,language translation needs proper algorithm to convert the sentences and by this it will help people we will able to change sentences from on language to another with a proper meaning.

Glossary

Transformer:-A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).

Encoder:- The encoder in the transformer consists of multiple encoder blocks. An input sentence goes through the encoder blocks, and the output of the last encoder block becomes the input features to the decoder.

Decoder:- The decoder also consists of multiple decoder blocks.

Self-Attention Models:-Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of a sequence.

2 Problem Definition

Scope

The project will accomplish the following objectives:

- Implementation of the transformer based attention model.
- Language Translation

Exclusions

We have 2 lakh training set, accuracy would be around 90percent.

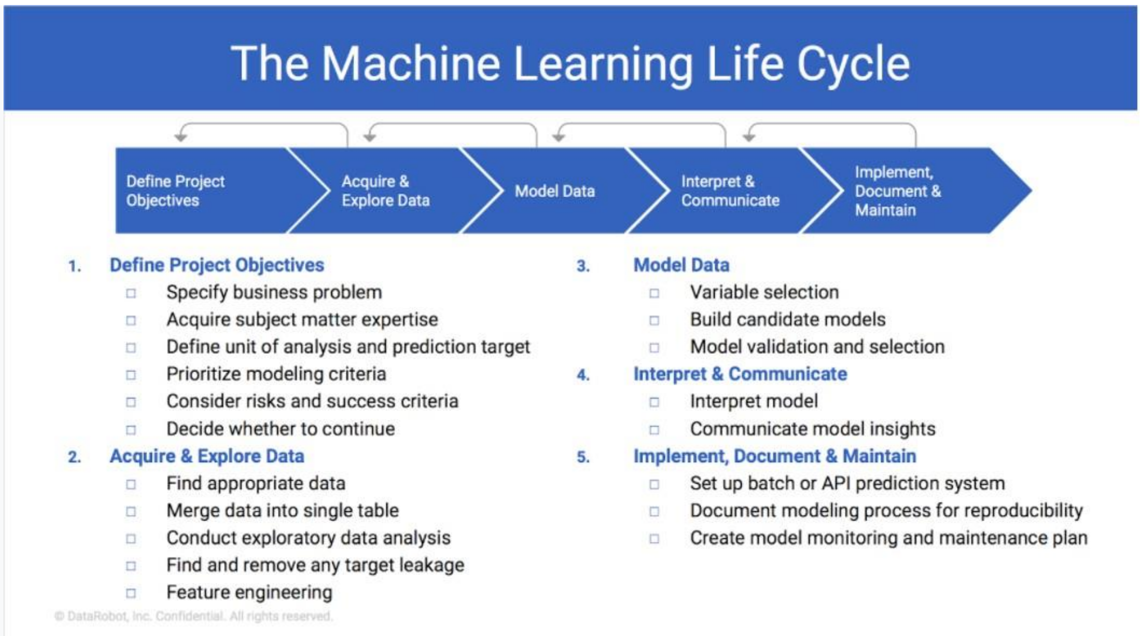
Assumptions

- The encoding component consists of 6 encoders stacked on top of each other.
- The decoding component is a stack of decoder of same number.
- The score is divided by 8 i.e square root of the dimension of the key vectors as it provides the most stable gradients.

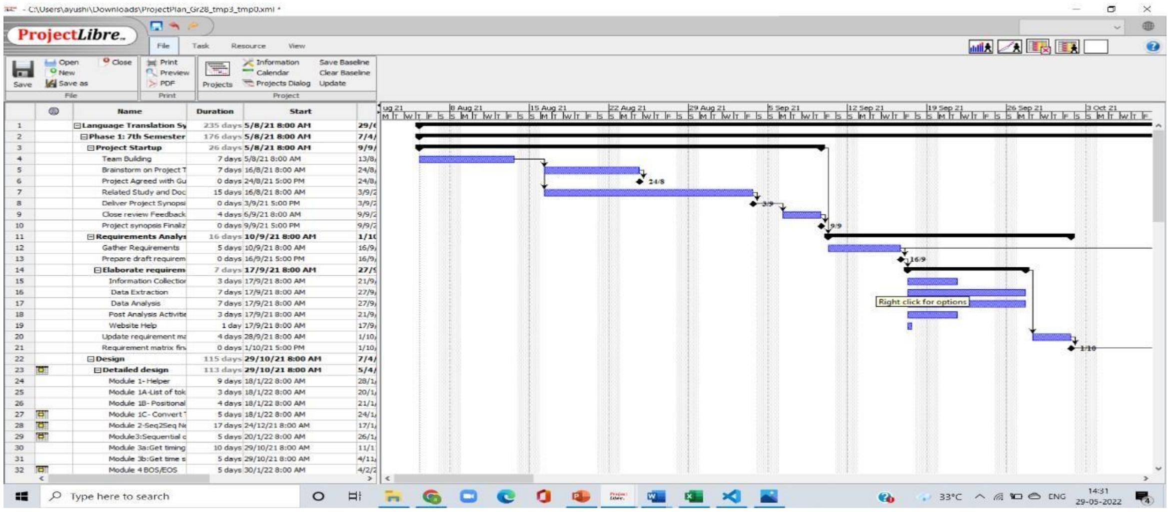
3 Project Planning

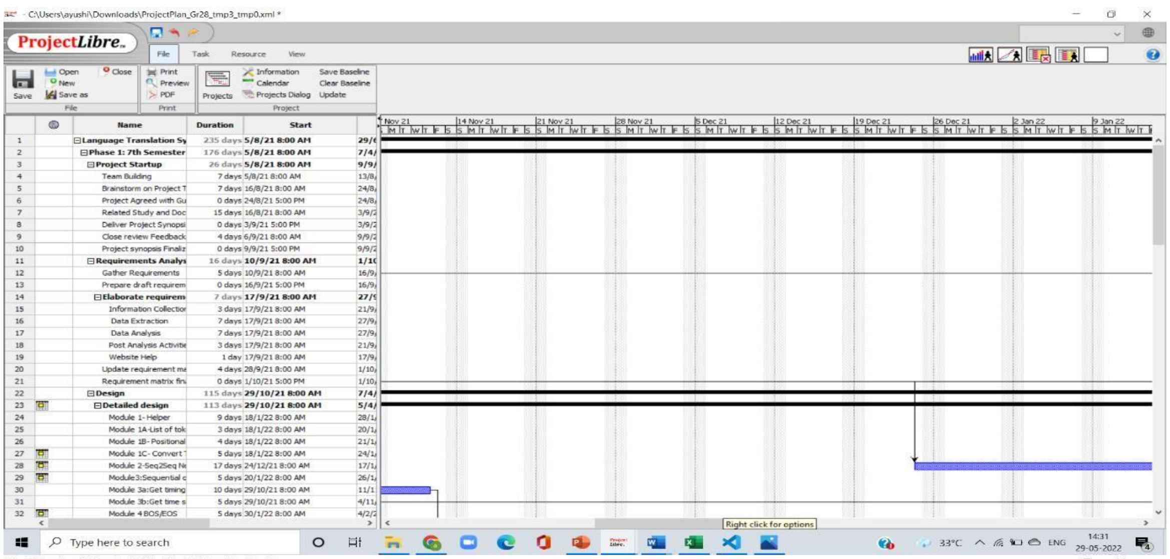
Software Life Cycle Model

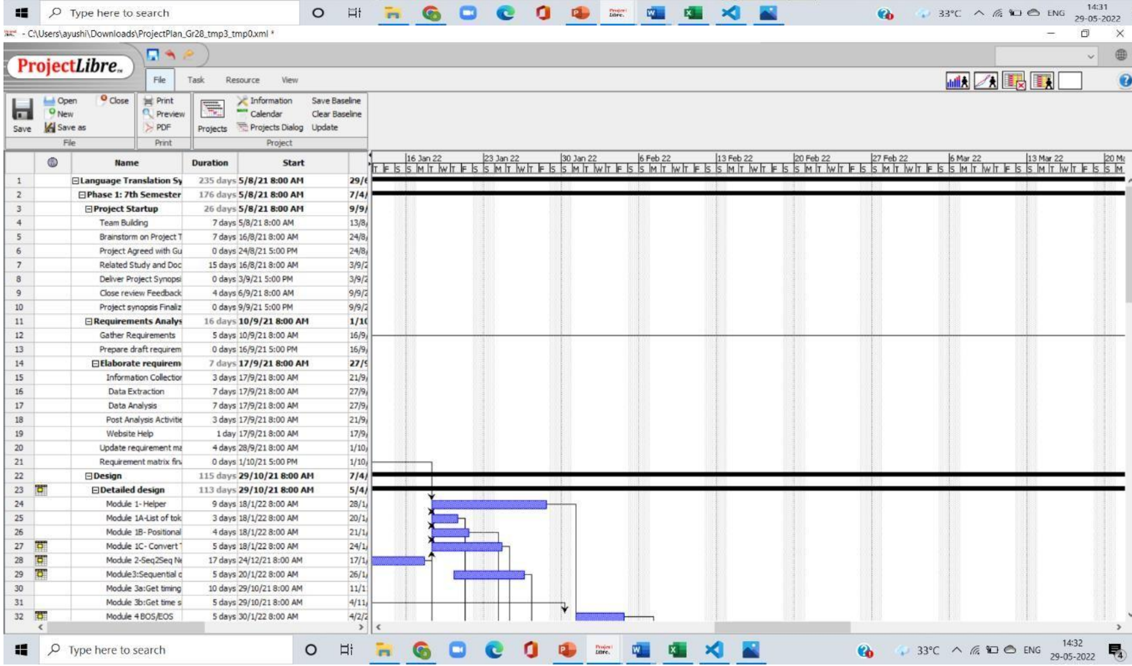
There are five major steps in the machine learning life cycle, all of which have equal importance and go in a specific order.



Scheduling







Cost Analysis

Our software is an application program. We are using Constructive Cost Model (COCOMO) for cost estimation.

Our project falls under the organic type with 42 kilo lines of code (KLOC).

We shall do the calculations using the basic model because the basic COCOMO is used in Organic mode by default.

Basic model:

The arithmetic formula of Basic COCOMO is:

a) Effort applied to the project: E = a$(KLOC)^b$ (in person-months)

b) Development time: D = c$(E)^d$ (in month)

where a, b, c, d are constraints for each category of software product.

Software Project	a	b	c	d
Organic	2.4	1.05	2.5	0.38
Semi-Detached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

so for our model a=2.4, b= 1.05, c= 2.5 and d=0.38.

Effort applied to the Project:

E = a$(KLOC)^b$

= 2.4 *$(42)^{1.05}$

= 2.4 * 50.63

= 121.512-person month

Development Time:

D = $c(E)^{d}$

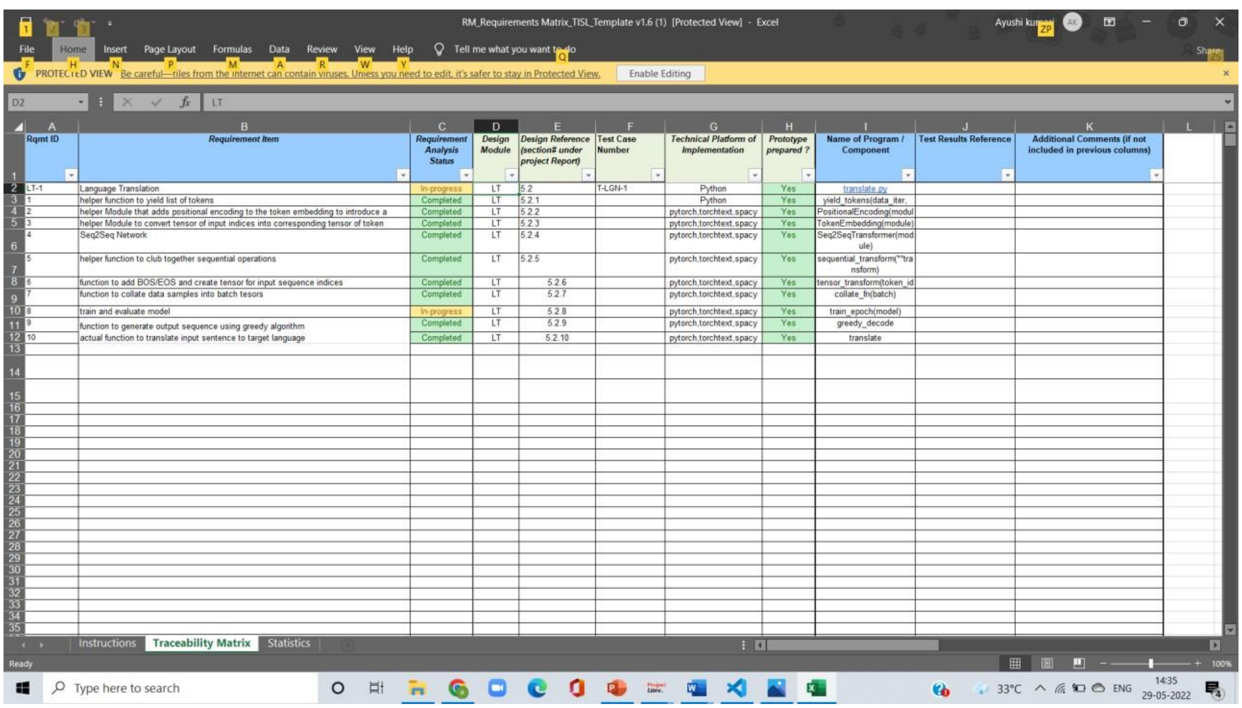
= $2.5 * (121.512)^{0.38}$

= $2.5 * 6.20$

= 15.5 months

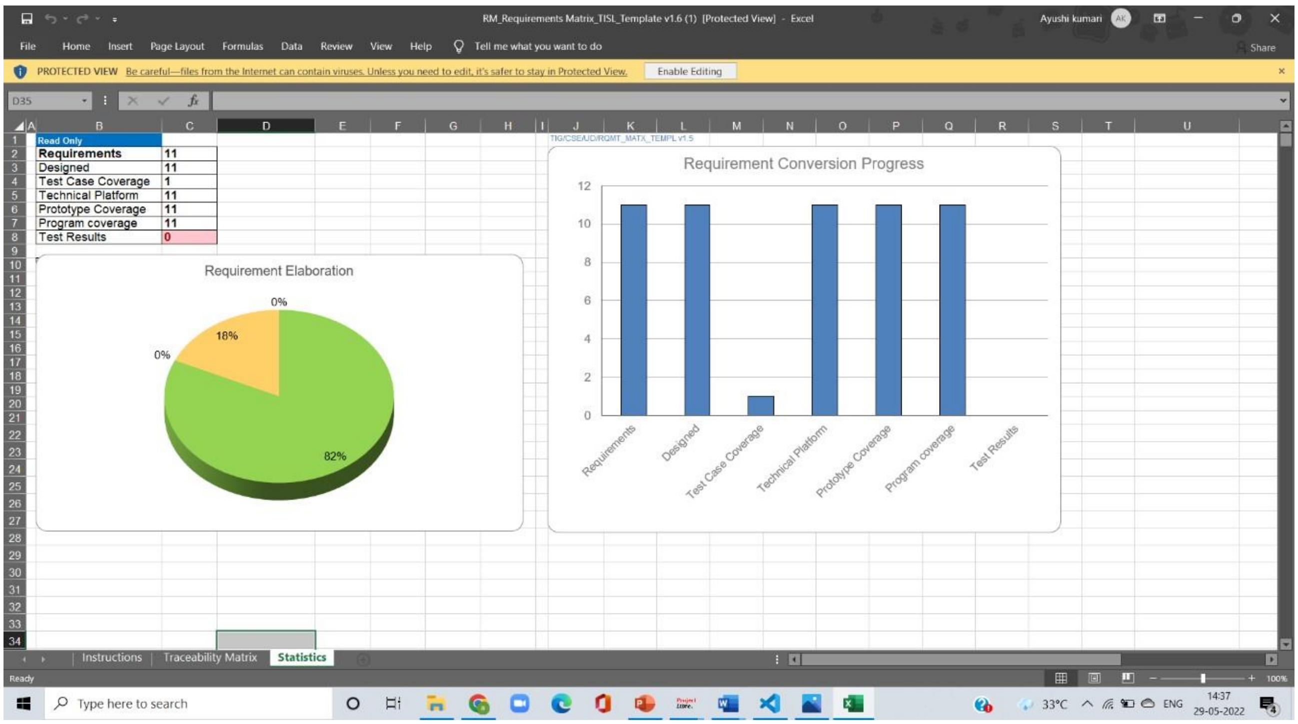
4 Requirement Analysis

Requirement Matrix



Rqmt ID	Requirement Item	Requirement Analysis Status	Design Module	Design Reference (section# under project Report)	Test Case Number	Technical Platform of Implementation	Prototype prepared ?	Name of Program / Component	Test Results Reference	Additional Comments (if not included in previous columns)
LT-1	Language Translation	In-progress	LT	5.2	T-LGN-1	Python	Yes	translate.py		
1	helper function to yield list of tokens	Completed	LT	5.2.1		Python	Yes	yield_tokens(data_iter,		
2	helper Module that adds positional encoding to the token embedding to introduce a	Completed	LT	5.2.2		pytorch,torchtext,spacy	Yes	PositionalEncoding(modul		
3	helper Module to convert tensor of input indices into corresponding tensor of token	Completed	LT	5.2.3		pytorch,torchtext,spacy	Yes	TokenEmbedding(module)		
4	Seq2Seq Network	Completed	LT	5.2.4		pytorch,torchtext,spacy	Yes	Seq2SeqTransformer(module)		
5	helper function to club together sequential operations	Completed	LT	5.2.5		pytorch,torchtext,spacy	Yes	sequential_transform(*transform)		
6	function to add BOS/EOS and create tensor for input sequence indices	Completed	LT	5.2.6		pytorch,torchtext,spacy	Yes	tensor_transform(token_id		
7	function to collate data samples into batch tesors	Completed	LT	5.2.7		pytorch,torchtext,spacy	Yes	collate_fn(batch)		
8	train and evaluate model	In-progress	LT	5.2.8		pytorch,torchtext,spacy	Yes	train_epoch(model)		
9	function to generate output sequence using greedy algorithm	Completed	LT	5.2.9		pytorch,torchtext,spacy	Yes	greedy_decode		
10	actual function to translate input sentence to target language	Completed	LT	5.2.10		pytorch,torchtext,spacy	Yes	translate		

Requirement Elaboration



5 Design

Technical Environment

Hardware:

20 gm RAM

256 SSD

10 GB Graphic Card

Operating System:

Ubuntu

Software:-

Text Editor

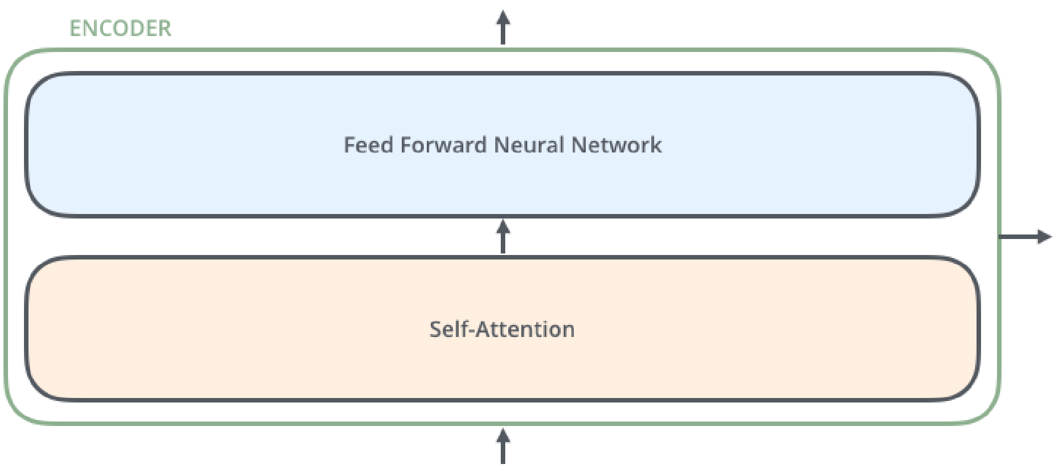
Spacy

TorchText

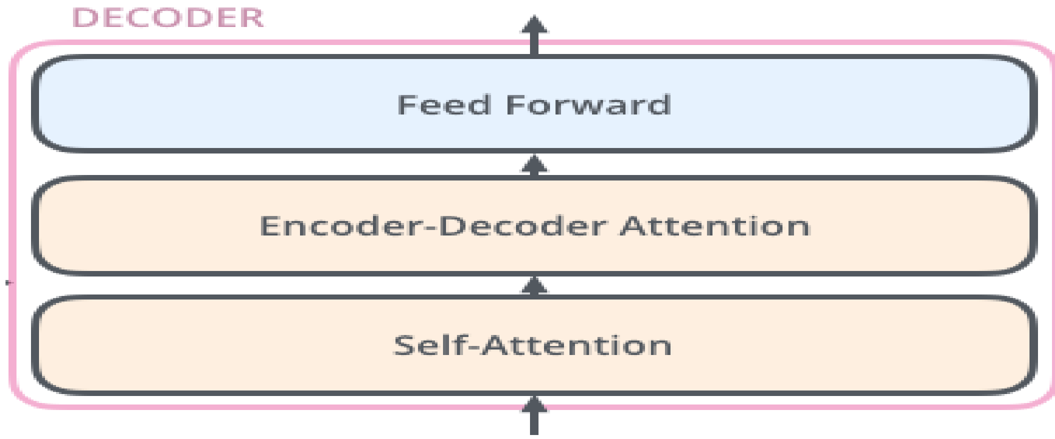
pyTorch

Detailed Design

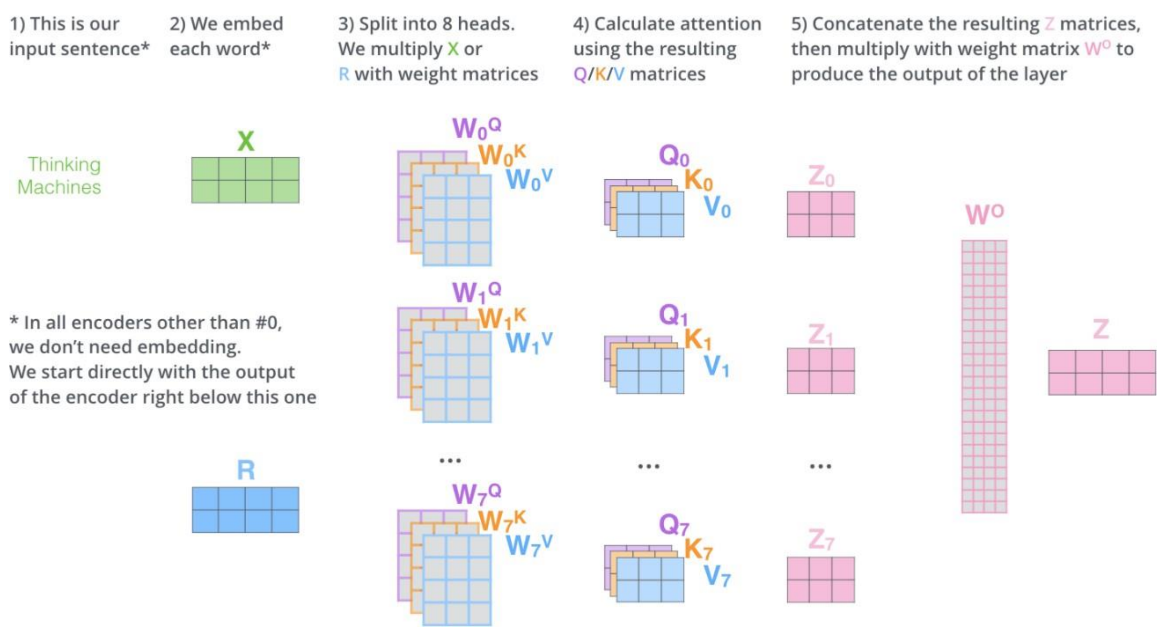
Encoder:-



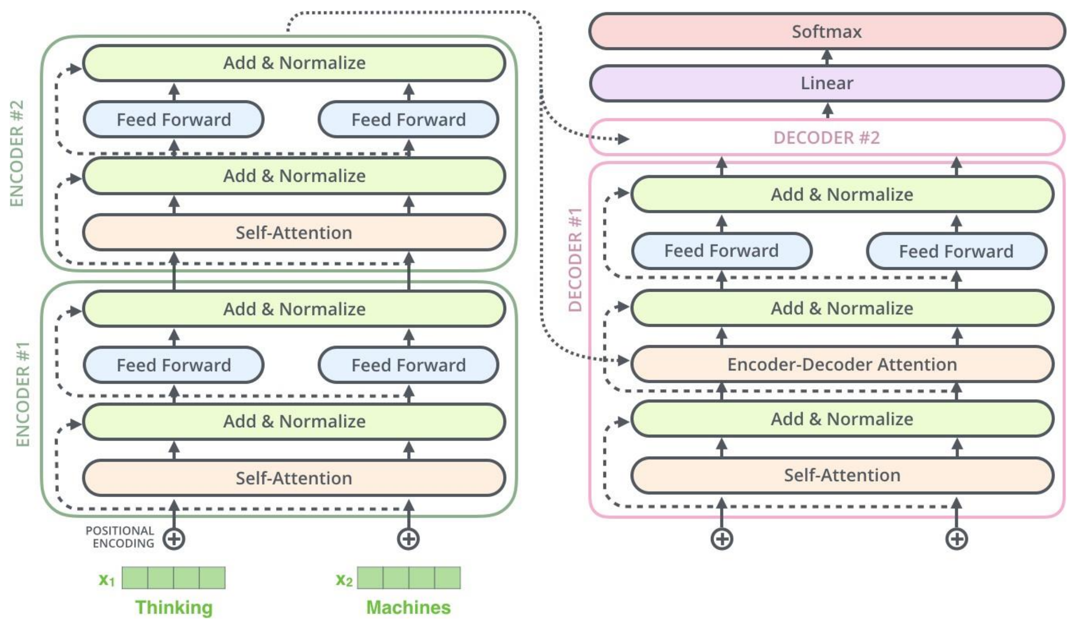
Decoder:-

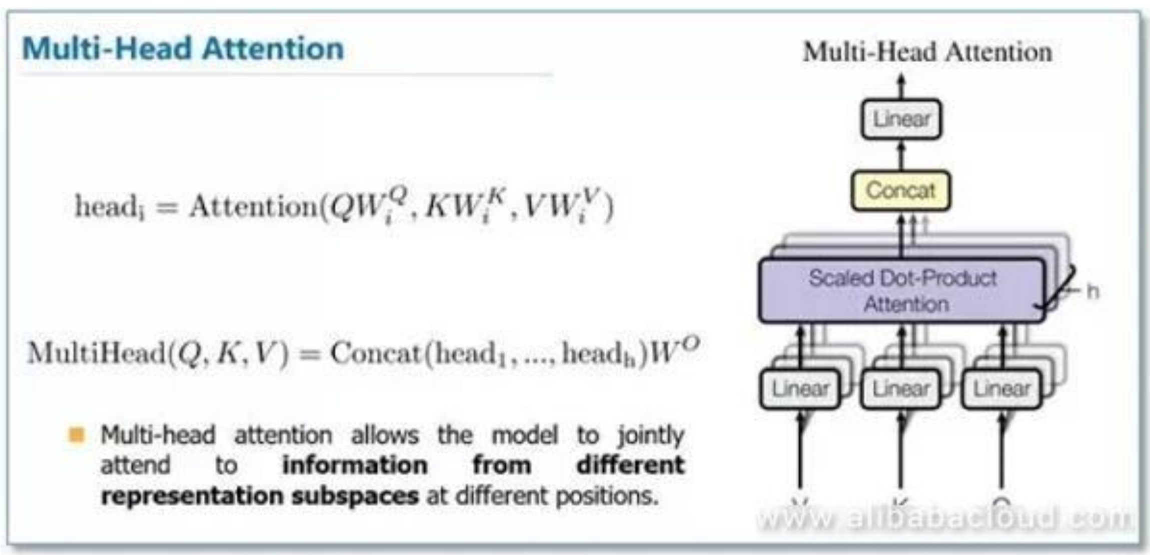


Multi-headed self-attention

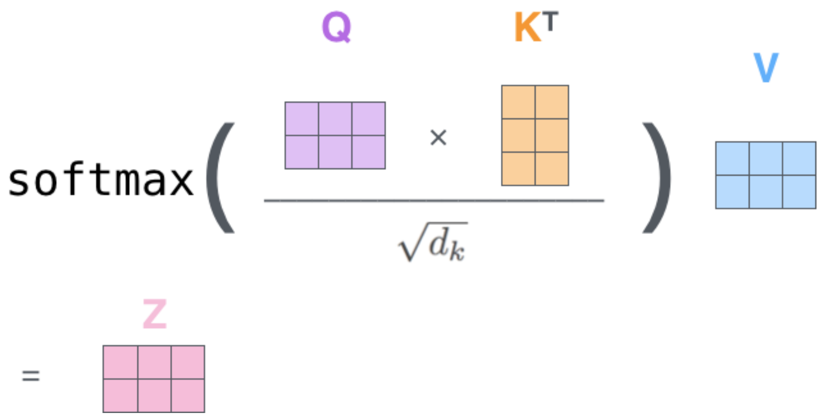


Overview of encoder and decoder

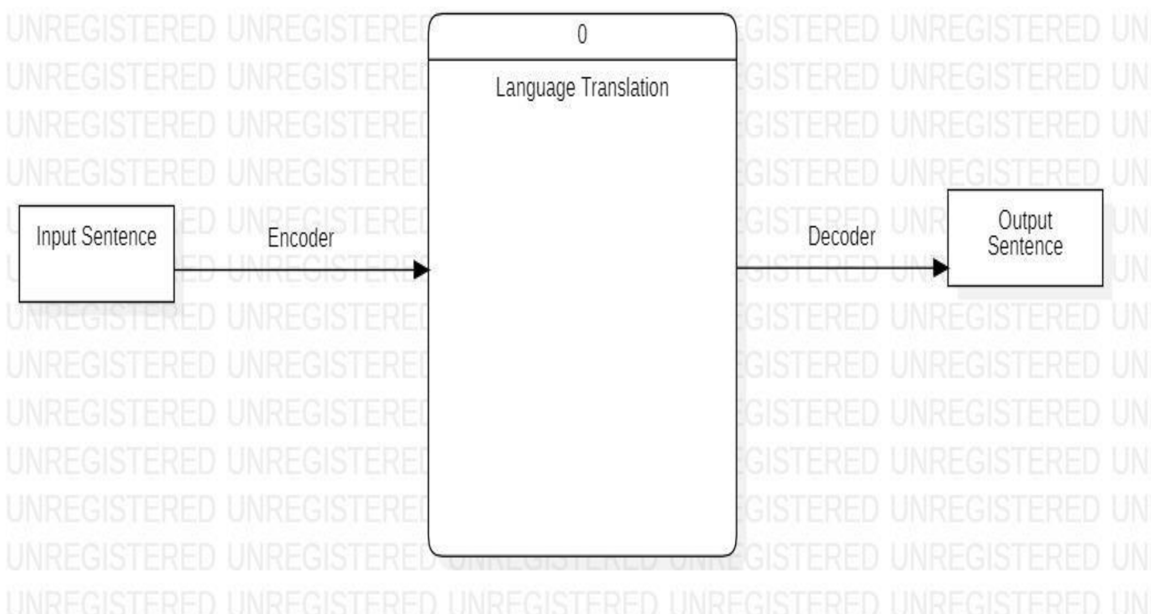




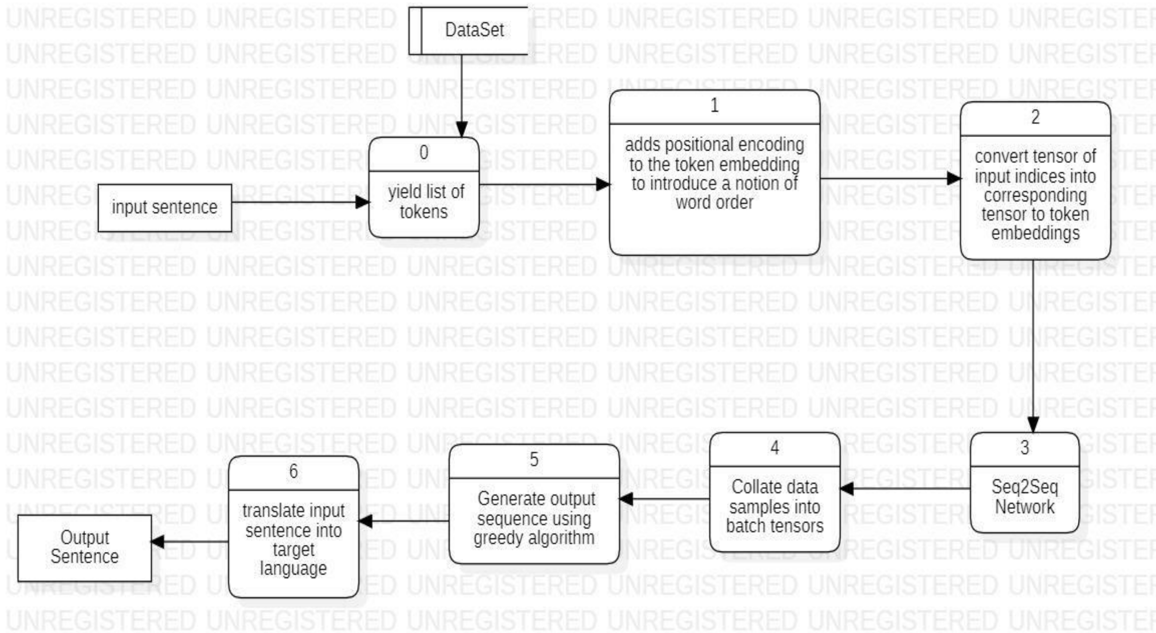
Calculation of SoftMax



DFD Level-0



DFD Level-1



6 Implementation

Implementation Details

a. Module to yield list of tokens

b. Module that adds positional encoding to the token embedding to introduce a notion of word order.

c. Module to convert tensor of input indices into corresponding tensor to token embeddings.

d. Seq2Seq Network

e. Collate data samples into batch tensors

f. Generate output sequence using greedy algorithm

g. Function to translate input sentence into target language

h. Source and target language text transforms to convert raw strings to tensors indices

System Installation Steps

We have installed four software for running the project:-

1. Python

   Commands :-

   Sudo apt update

   Sudo apt install software-properties-common

   Sudo apt install python3.8

2. Spacy

   Commands:-

   Pip install -U pip setuptools wheel

   Pip install -U spacy

   Pip -m spacy download en core wen sm

3. Torchtext

   Commands:-

   Pip install torchtext

4. PyTorch

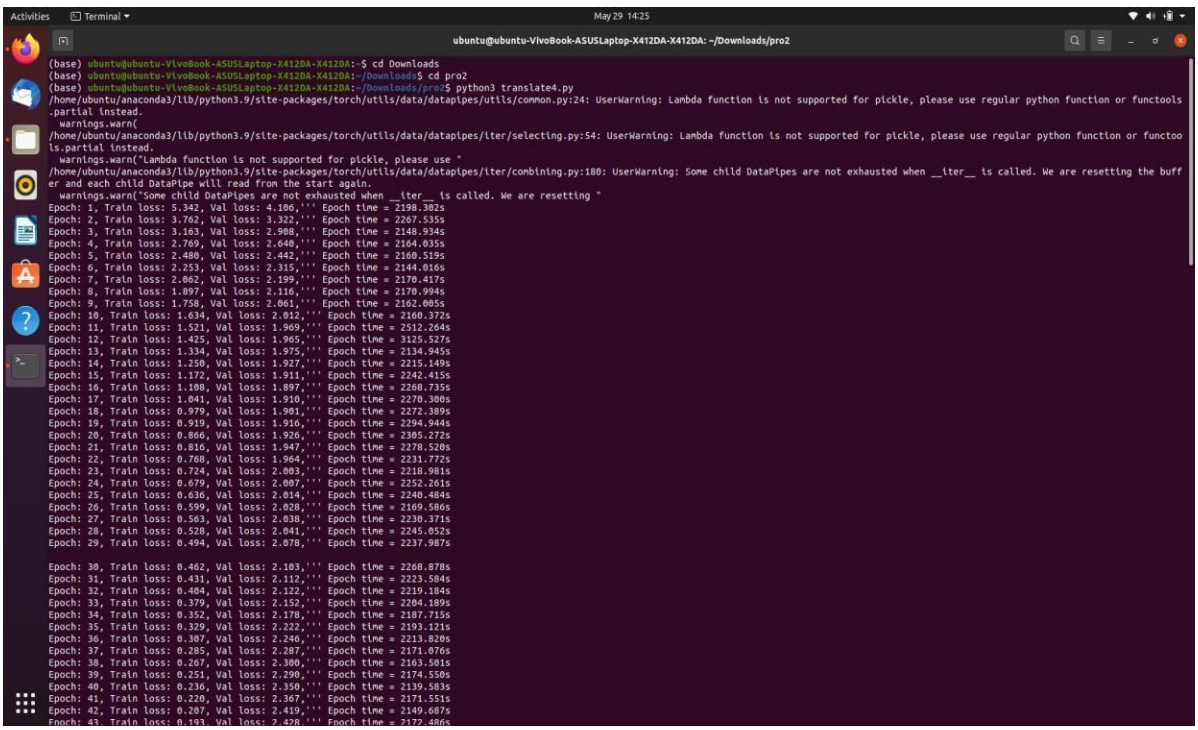
Commands:-

Sudo apt install python3-pip

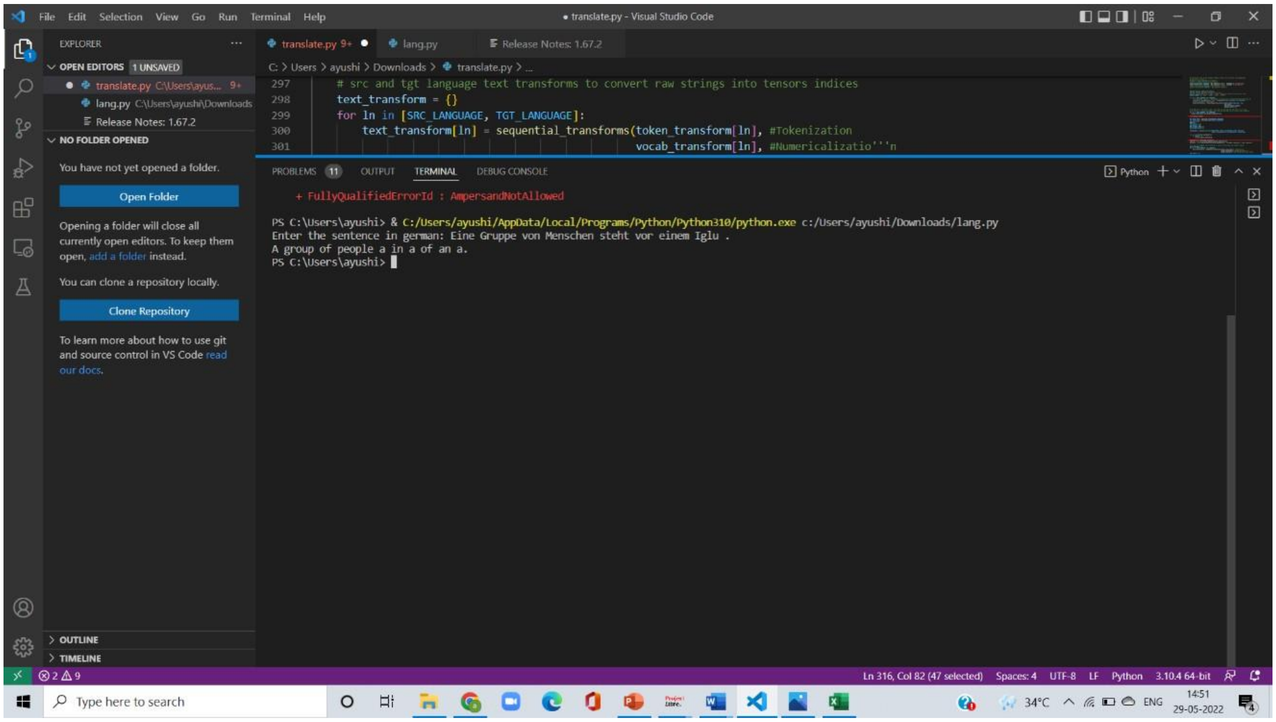
Pip3 install torch==1.9+cpu torchvision==0.10.1+cpu-f https://downloadpytorch.org/whl/torch_stable.html

System Usage Instructions

Command to run the application:-



7 Test Results and Analysis

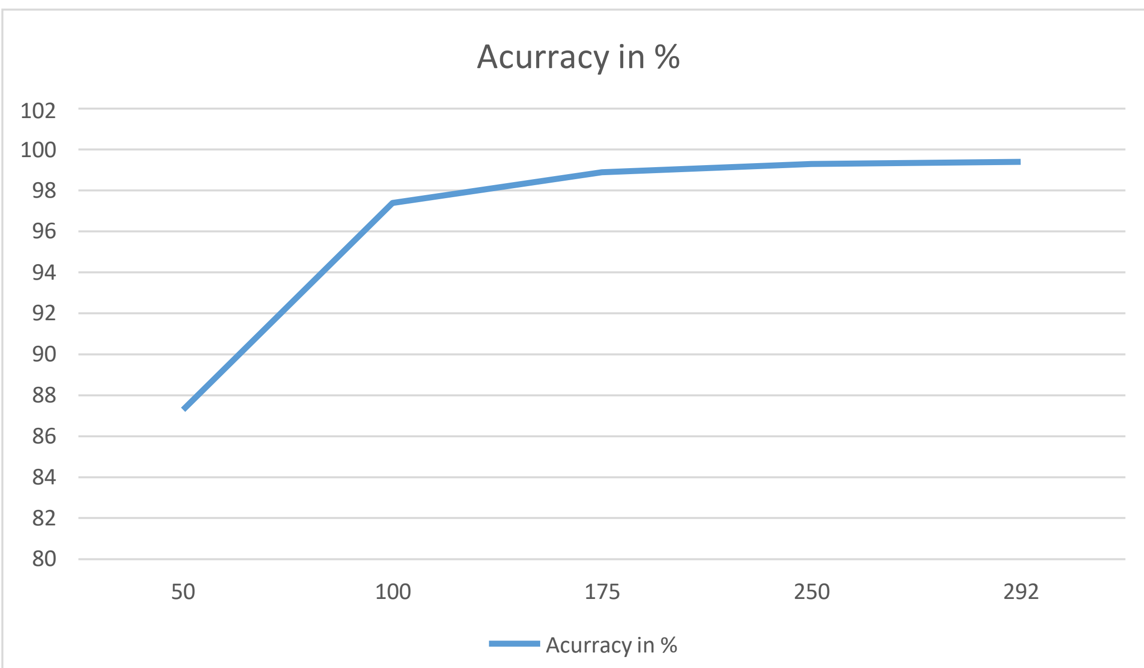


Serial No.	Test Cases	Output Result	Expected Result	Status
1.	Eine Gruppe von Menschen stehtvoreinem Iglu.	A group of people a in a of an a.	A group ofpeople stand in front of an igloo.	40 percent
2.	Es istunserAbschlussprojekt	It's our final a project	It’s our final Year project	50 percent
3.	Unser ProjektnameistSprachübersetzung	Our project a	Our project name is	40 percent

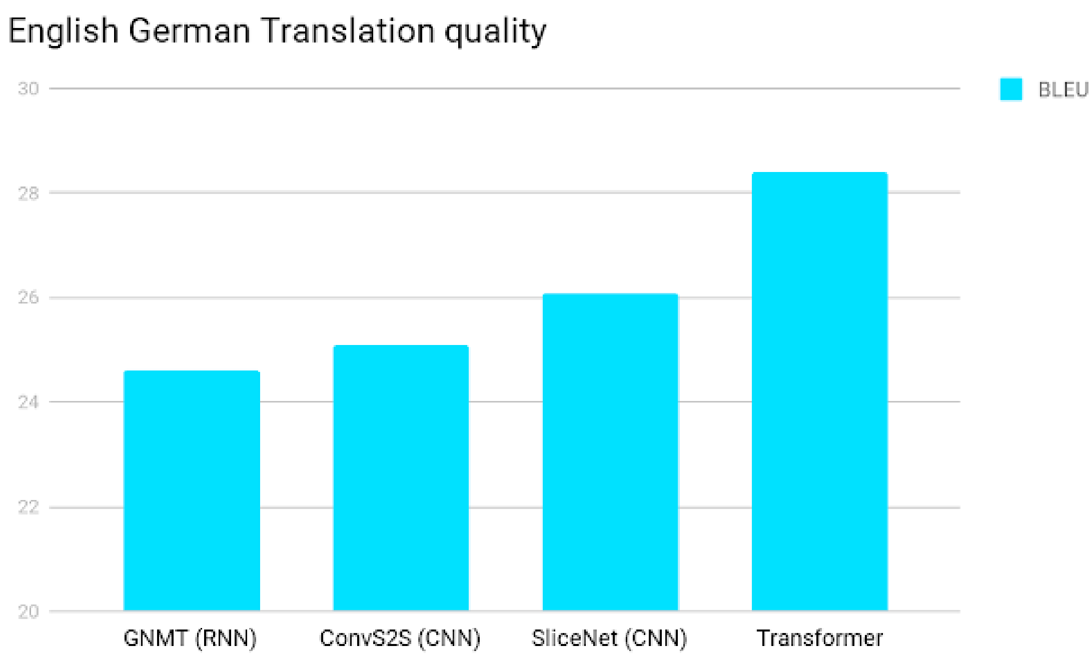
		Language a	language translation	

Performance Metrics:

SL.NO.	Epoch size	Training loss	Validation Loss	Accuracy
1	50	0.127	2.521	87.3%
2	100	0.026	3.193	97.4%
3	175	0.011	3.541	98.9%
4	250	0.007	3.733	99.3%
5	292	0.006	3.792	99.4%



Comparative study:



Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Analysis:-

The test cases is almost 40%acccuarte.

It depends upon how many epochs we are running .More epochs,

more learning will be there and more accurate result will come.

8 Conclusion

Project Benefits

The ability to communicate with one another is a fundamental part of being human. There are nearly 7,000 different languages worldwide. As our world becomes increasingly connected, language translation provides a critical cultural and economic bridge between people from different countries and ethnic groups. Some of the more obvious use-cases include:

- business: international trade, investment, contracts, finance
- commerce: travel, purchase of foreign goods and services, customer support
- media: accessing information via search, sharing information via social networks, localization of content and advertising
- education: sharing of ideas, collaboration, translation of research papers
- government: foreign relations, negotiation

To meet these needs, technology companies are investing heavily in machine translation. This investment and recent advancements in deep learning have yielded major improvements in translation quality. According to Google, switching to deep learning produced a 60% increase in translation accuracy compared to the phrase-based approach previously used in Google Translate. Today, Google and Microsoft can translate over 100 different languages and are approaching human-level accuracy for many of them.

Future Scope for improvements

As transformer based model can be used in sequential based data it can be used in protein sequence analysis and in DNA sequence analysis

9 References / Bibliography

1. Alammar, Jay. "The Illustrated Transformer – Jay Alammar – Visualizing Machine Learning One Concept at a Time." *Jay Alammar – Visualizing Machine Learning One Concept at a Time.*, https://jalammar.github.io/illustrated-transformer/. Accessed 2 June 2022.

2. Gaffar, Syed Abdul. "Language Translation with Transformer in Python!" *Analytics Vidhya*, 12 June 2021, https://www.analyticsvidhya.com/blog/2021/06/language-translation-with-transformer-in-python/.

3. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

4. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

5. Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. Deep recurrent models with fast-forward connections for neural machine translation. CoRR, abs/1606.04199, 2016