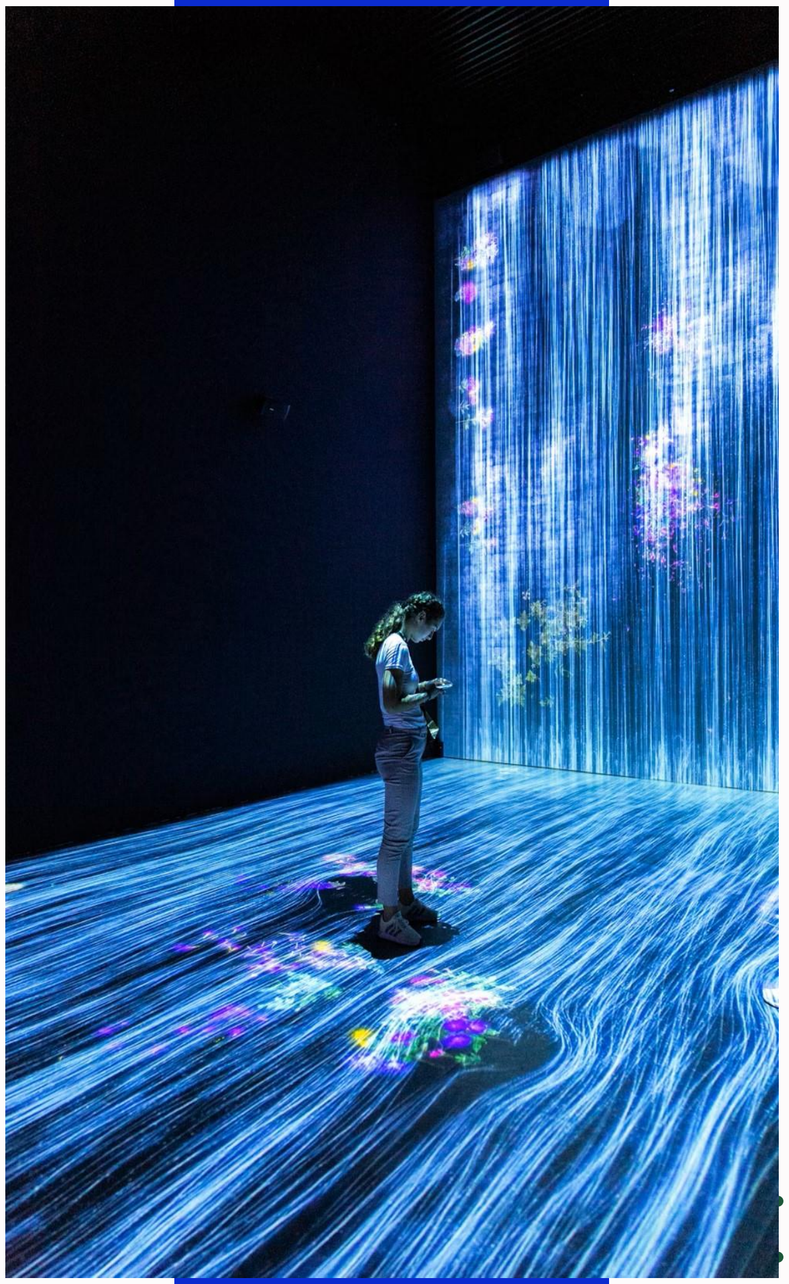
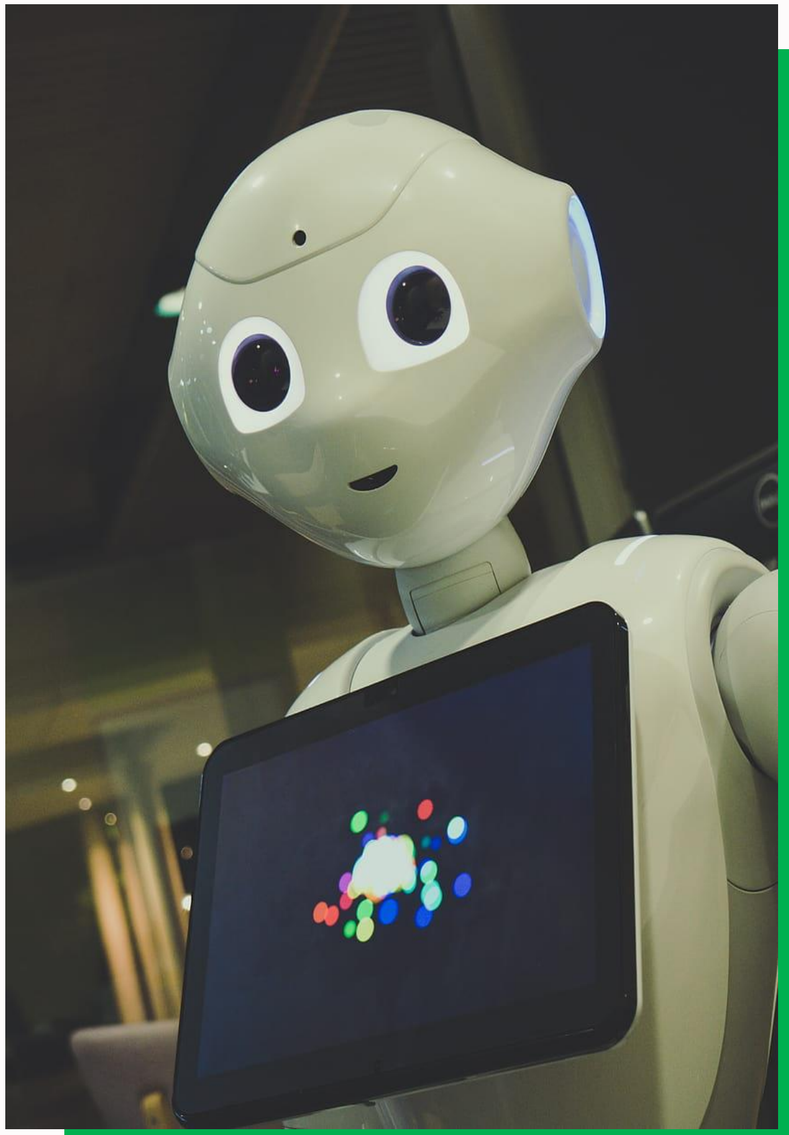
NEURAL MACHINE TRANSLATION



Content



Overview

Introduce the project. Provide a quick background and rationale. Briefly share its overall scope as well as expected outcomes.

01

A high-quality, diverse, and representative dataset is essential for training a neural machine translation (NMT) model effectively. The key aspects of the dataset's role in machine translation.

02

Long Short-Term Memory (LSTM) Networks: LSTM is a type of RNN that is designed to handle the vanishing gradient problem that can occur in standard RNNs. It does this by introducing three gating mechanisms that control the flow of information through the network: the input gate, the forget gate, and the output gate.

03

NMT models are trained on parallel corpora, where each sentence in the source language has a corresponding translation in the target language. Training involves optimizing the model parameters to minimize the difference between predicted and actual translations.

04

The future of translation lies in the collaboration between human intelligence and AI-powered machine translation. Human translators excel in creative thinking and adapting translations for specific audiences, while AI is ideal for automating repetitive tasks.

Our Team

Project Guide:

Prof.Koushik De

Team Members:

Name	University Roll No
▪ Soujit Das	14871022031
▪ Arkya Paul	14871022004
▪ Mainak Sarkar	14871022037
▪ Shouvik Saha	14871022006
▪ Sahil Raj	14871022052

INTRODUCTION

It is an information system where input data passes through several interconnected nodes to generate an output. Neural machine translation software uses neural networks to work with enormous datasets. Each node makes one attributed change of source text to target text until the output node gives the final result.

Motivation

Clarify the project's overall motivations.

A journey that was ignited by the power of research papers. We live in a world where information is at our fingertips, and the pursuit of knowledge has never been more accessible. It was through the pages of research papers that I found a profound source of motivation that has shaped my perspective and fueled our aspirations for this amazing project.

This experience sparked a fire within us an insatiable thirst for knowledge and a desire to contribute to the collective tapestry of human understanding. The research paper became more than just a document; it became a source of inspiration that fueled my determination to delve deeper, question further, and seek innovative solutions.

And a big thanks to our project guide who motivated us to take this challenge,And helped us a lot through our whole journey.

DATASET USED

Introduce the main dataset which is trained through the neural network

English to Bengali

An English to Bengali dataset for neural machine translation (NMT) is a collection of paired sentences in English and Bengali languages. Each pair consists of an English sentence and its corresponding translation in Bengali. This dataset is crucial for training and evaluating machine translation models that aim to automatically convert text from English to Bengali.

- **An overview of used dataset**

What are you going to be when you grow up?	তুই বড়ো হয়ে কী হবি?
What time does the train for Boston leave?	বস্টনের ট্রেনটা কটার সময় ছাড়ে?
Why were you absent from school yesterday?	তুমি কাল বিদ্যালয়ে অনুপস্থিত ছিলে কেন?
You never mentioned you were Tom's friend.	আপনি কখনই উল্লেখ করেননি যে আপনি টমের বন্ধু ছিলেন।
You're going to have to come with us, Tom.	টম, তোমাকে আমাদের সাথে আসতে হবে।
You're in no position to make any demands.	আপনি কোনো কিছু দাবী করার মতন জায়গায় নেই।
You're supposed to be working, aren't you?	আপনার কাজ করার কথা, তাই না?
You've known Tom a lot longer than I have.	আপনি টমকে আমার থেকে অনেক বেশী দিন ধরে চেনেন।
Pass me the salt, please. "Here you are."	অনুগ্রহ করে নুনটা এগিয়ে দেবেন। "এই নিন।"
Pass me the salt, please. "Here you are."	অনুগ্রহ করে নুনটা এগিয়ে দেবেন। "এই নাও।"
Are tickets for the concert available here?	সঙ্গীতানুষ্ঠানের জন্য টিকিট কি পাওয়া যাচ্ছে?
Besides teaching English, he writes novels.	ইংরেজি শেখানোর পাশাপাশি তিনি উপন্যাস লেখেন।
Could you please repeat what you just said?	আপনি যেটা বললেন সেটা কি দয়া করে আর একবার বলতে পারবেন?
Didn't you know Tom used to live in Boston?	আপনি জানতেন না যে টম বোস্টনে থাকে?
Give me a couple of minutes alone with Tom.	আমাকে একলা টমের সঙ্গে দু মিনিট সময় দাও।
How could I be a robot? Robots can't dream.	আমি কীভাবে রোবট হতে পারি? রোবট স্বপ্ন দেখতে পারে না।
I can't seem to focus on work at all today.	আমি আজকে কাজে মনযোগ দিতে পারছি না।
I should've hit Tom while I had the chance.	যখন আমার কাছে সুযোগ ছিলো তখনই আমার টমকে আঘাত করা উচিত ছিল।
I think you'll probably be able to do that.	আমি মনে করি আপনি সম্ভবত এটি করতে সক্ষম হবেন।
I used to go to Australia when I was a kid.	আমি ছোটবেলায় অস্ট্রেলিয়া যেতাম।
Is she singing a song or playing the piano?	সে কি গান গাইছে না কি পিয়ানো বাজাচ্ছে?
Nothing that Tom does surprises me anymore.	টমের কোন কাজই এখন আমাকে অবাক করে না।
The Battle of Gettysburg lasted three days.	গেটিসবার্গের যুদ্ধ তিন দিন স্থায়ী হয়েছিল।
Tom is the only one who doesn't seem happy.	একমাত্র টমকেই খুশি বলে মনে হচ্ছে না।
Tom said he knew who Mary's ex-husband was.	টম বলেছিল যে সে জানে মেরির প্রাক্তন স্বামী কে।
Tom said he knew who Mary's ex-husband was.	টম বলেছিল মেরির প্রাক্তন স্বামী কে তা সে জানে।
Tom sat alone on the porch smoking a cigar.	টম একা একা উঠোনে বসে সিগার খাচ্ছিলো।

Hardware & Software

The choice of hardware and software in a project is a critical decision that influences the project's performance, scalability, and overall success. Here's a brief summary of the key considerations for both hardware and software in this project

- ✓ Microsoft Excel
- ✓ Google Chrome(browser)

- Anaconda Navigator- v2.4.2
- Jupyter Notebook- v6.5.4
- Python v3.1

- Operating System- Windows 10 Home, 64-bit operating system

- Processor- 11th Gen Intel(R) Core (TM) i3-1115G4 3.00GHz
- Installed RAM- 4.00 GB

Methodology

01

Data Collection:

This is a collection of texts in the source language aligned with their translations in the target language. The quality and size of this dataset significantly impact the performance of the NMT model.

02

Data Preprocessing:

•**Tokenization:** Break down sentences into smaller units, usually words or subword units.
•**Vocabulary creation:** Build a vocabulary for both the source and target languages based on the tokenized data.

03

Model Architecture:

•**Encoder-Decoder Architecture:** The model typically consists of an encoder and a decoder. The encoder processes the input sequence, and the decoder generates the output sequence.
•**Attention Mechanism:** To allow the model to focus on different parts of the input sequence when generating each part of the output sequence.

04

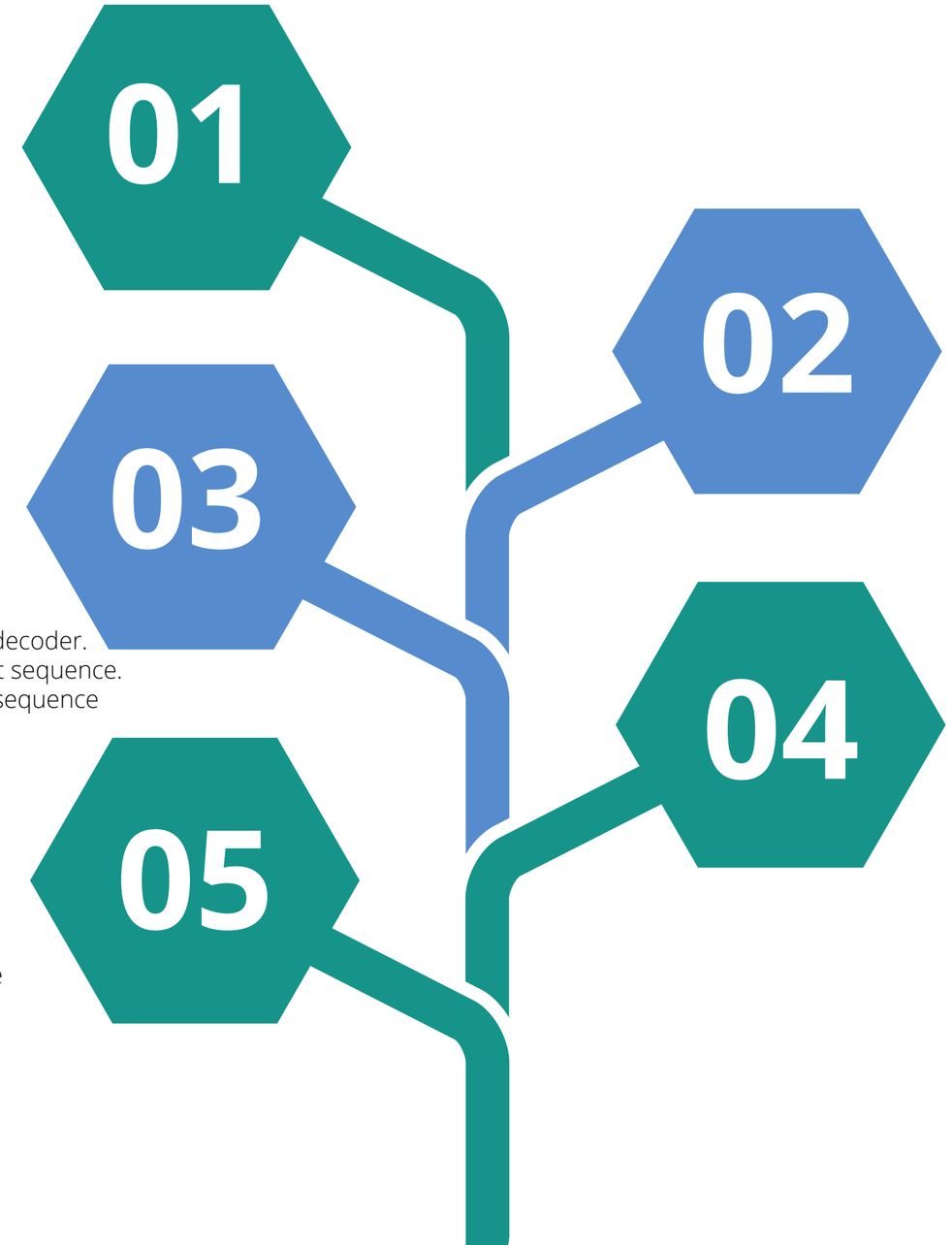
Embeddings:

•**Word embeddings:** Represent words as continuous vectors to capture semantic relationships between words.
•**Positional embeddings:** Incorporate the order or position of words in a sequence.

05

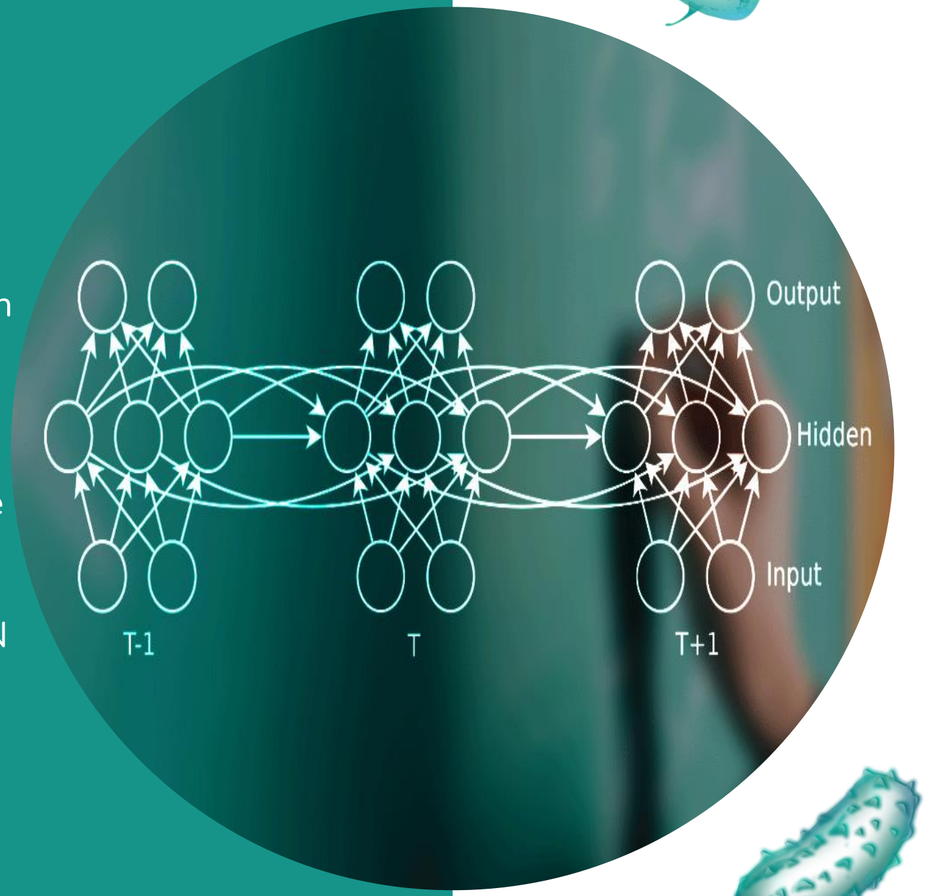
Training:

•**Loss function:** Use a suitable loss function, such as cross-entropy loss, to measure the difference between the predicted and actual translations.
•**Backpropagation:** Optimize the model parameters using gradient descent or a variant to minimize the loss.



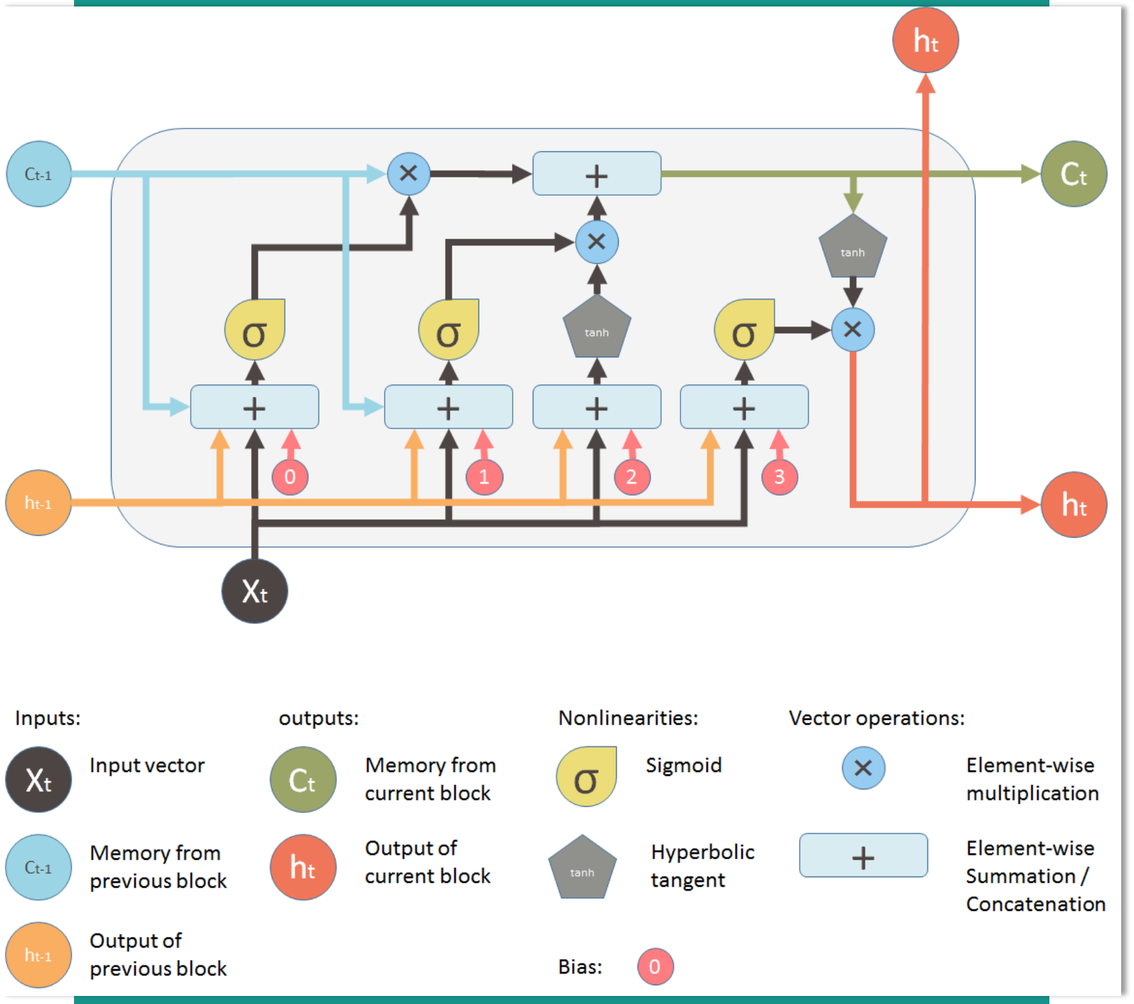
Recurrent Neural Network

- Recurrent Neural Network(RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is its **Hidden state**, which remembers some information about a sequence. The state is also referred to as *Memory State* since it remembers the previous input to the network. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks.



LSTM

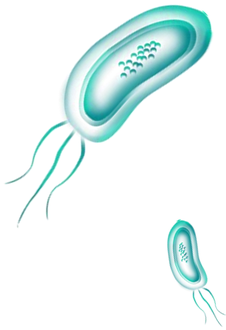
(Long Short-Term Memory)

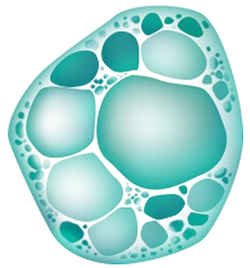


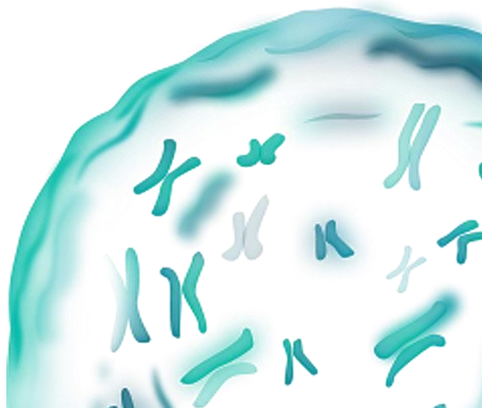
About

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that addresses the vanishing gradient problem, a common issue in traditional RNNs. The vanishing gradient problem occurs when training deep networks, and it makes it difficult for the model to learn dependencies between distant time steps. LSTM was introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 to overcome some challenge.

LSTMs are a pivotal development in the evolution of recurrent neural networks. They have proven to be effective in capturing long-term dependencies in sequential data, making them a valuable tool in various machine learning applications, especially those involving natural language understanding and generation.







Experimental Result

English Text

"Dynamite was invented by Alfred Nobel in 1857."

Example 01

Translated to Bengali

১৮৫৭ সালে আলফ্রেড নোবেল ডিনামাইট আবিষ্কার করেন।.

English Text

"Wikipedia is the best encyclopedia on the Internet."

Example 02

Translated to Bengali

উইকিপিডিয়া হলো ইন্টারনেটে সবচেয়ে সেরা বিশ্বকোষ।

English Text

"I thought that doing this would be easy, but we've been working all day and we're still not finished. "

Example 03

Translated to Bengali

আমি ভেবেছিলাম এটা করা সহজ হবে, কিন্তু আমরা সারাদিন ধরে কাজ করেছি আর এখনো শেষ করে উঠতে পারিনি। .

English Text

"I heard they found a skeleton buried in the basement of a house on Park Street. "

Example 04

Translated to Bengali

আমি শুনলাম যে তাঁরা পার্ক স্ট্রিটের একটা বাড়ির ভিতের মধ্যে থেকে একটা কঙ্কাল পেয়েছেন। .

Future goals

01

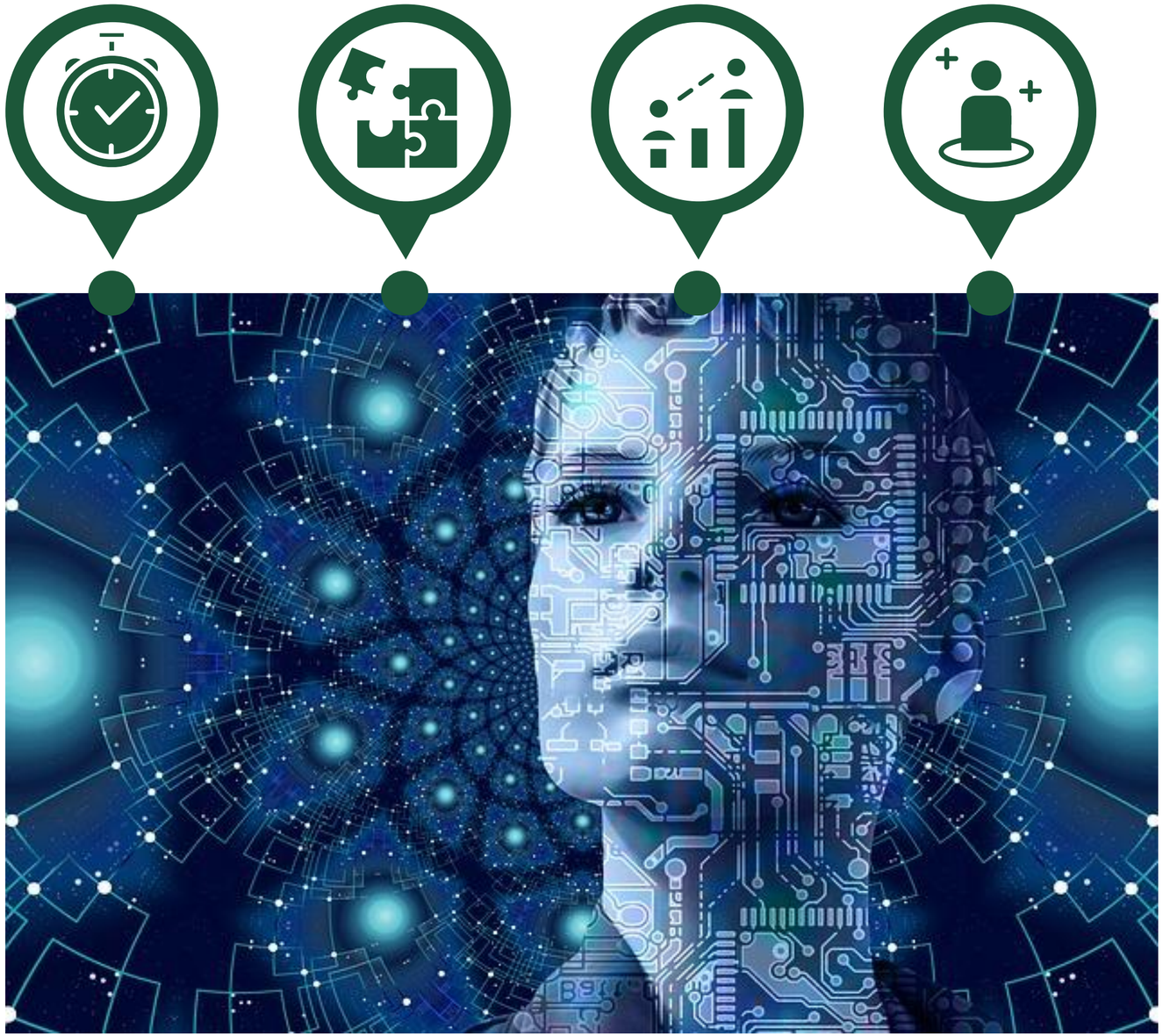
Improved Translation Quality:
•Continuously strive to enhance the translation quality by refining the neural network architecture, exploring advanced algorithms, and leveraging state-of-the-art pre-trained language models.

02

Multilingual Capability:
•Extend language support to include additional languages, especially those that are less commonly spoken, to make the machine translation system more inclusive and globally applicable.

03

Domain-Specific Adaptation:
•Develop mechanisms to adapt the machine translation model to specific domains, such as medical, legal, technical, or scientific, to improve the accuracy and relevance of translations in specialized fields.



Conclusions

In conclusion, the Neural Machine Translation (NMT) project represents a significant leap forward in breaking down language barriers and fostering global communication. The journey of this project has been marked by innovation, challenges, and continuous learning, with a commitment to advancing the capabilities of machine translation. As we reflect on the achievements and look toward the future, several key points emerge

Conclusion 01

Neural Machine Translation project stands as a testament to the transformative power of technology in facilitating cross-cultural communication. The journey has been marked by achievements, lessons learned, and a steadfast commitment to pushing the boundaries of what is possible in the realm of language translation. As the project moves forward, it does so with a vision of creating a more connected and inclusive world through the power of neural machine translation.

Conclusion 02

The project recognizes the dynamic nature of language and the importance of adaptability. Plans for continuous training and updating will ensure that the machine translation system remains relevant and effective in the face of evolving linguistic patterns and emerging vocabulary.

Conclusion 03

Looking ahead, the NMT project is poised for further growth and innovation. Future goals encompass enhanced translation quality, domain-specific adaptation, real-time capabilities, and a commitment to ethical and inclusive translation practices.

THANK YOU