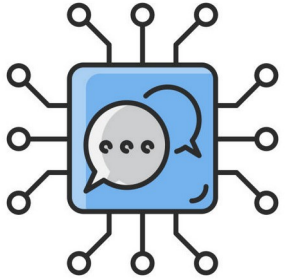
Seq-2-Seq Models

Tushar B. Kute,
http://tusharkute.com



Seq2Seq

- Seq2Seq model or Sequence-to-Sequence model, is a machine learning architecture designed for tasks involving sequential data.
- It takes an input sequence, processes it, and generates an output sequence.
- The architecture consists of two fundamental components: an encoder and a decoder.
- Seq2Seq models have significantly improved the quality of machine translation systems making them an important technology.

Seq2Seq

- The seq2Seq model is a kind of machine learning model that takes sequential data as input and generates also sequential data as output.
- Before the arrival of Seq2Seq models, the machine translation systems relied on statistical methods and phrase-based approaches.
- The most popular approach was the use of phrase-based statistical machine translation (SMT) systems. That was not able to handle long-distance dependencies and capture global context.

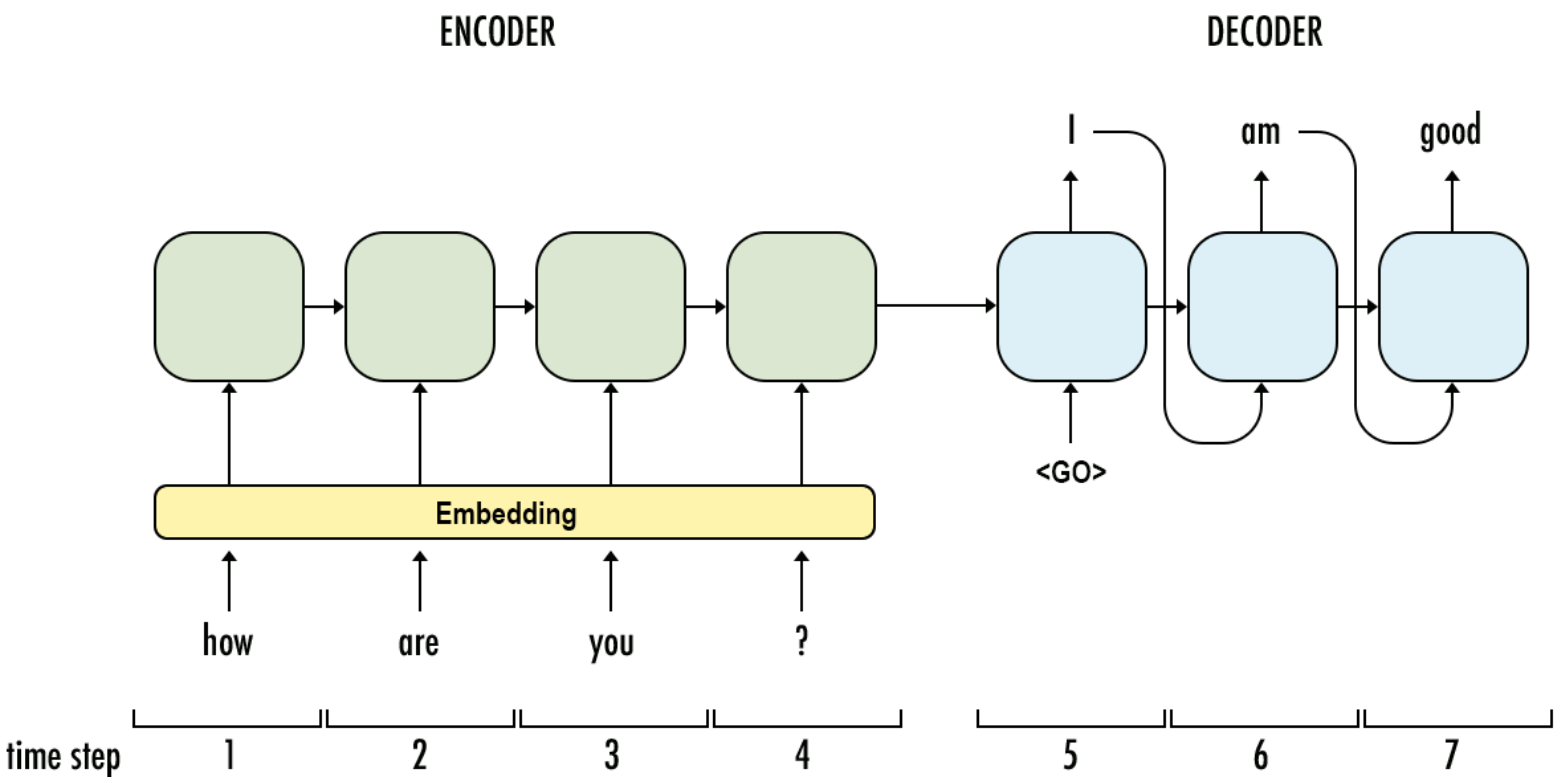
Seq2Seq

- Seq2Seq models addressed the issues by leveraging the power of neural networks, especially recurrent neural networks (RNN).
- The concept of seq2seq model was introduced in the paper titled "*Sequence to Sequence Learning with Neural Networks*" by Google.
- The architecture discussed in this research paper is fundamental framework for natural language processing tasks.
- The seq2seq models are encoder-decoder models. The encoder processes the input sequence and transforms it into a fixed-size hidden representation. The decoder uses the hidden representation to generate output sequence.

Seq2Seq

- A Sequence-to-Sequence (Seq2Seq) model is a type of neural network architecture used for tasks where both the input and the output are sequences. It is widely used for tasks such as:
  - Machine Translation: Converting sentences in one language to another (e.g., English to French).
  - Text Summarization: Summarizing long articles into shorter versions.
  - Speech Recognition: Converting audio signals into text.
  - Text Generation: Generating a sequence of text based on input context (e.g., chatbot systems).
  - Image Captioning: Generating descriptions for images.

Seq2Seq: Components



Seq2Seq: Components

- Encoder:
  - The encoder takes the input sequence and processes it into a fixed-size context vector (a single representation) that captures the essential information from the input sequence.
  - The encoder is typically implemented using a Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), or Gated Recurrent Unit (GRU) layers.
  - The output of the encoder is a "context vector," which is a compressed version of the input sequence.

Seq2Seq: Components

- Decoder:
  - The decoder takes this context vector as its input and generates the output sequence step by step.
  - The decoder also typically uses RNNs, LSTMs, or GRUs.
  - The decoder generates the output sequence one token at a time, using the context vector and its previous outputs as input for generating the next token.

Seq2Seq: Working

- Input Sequence:
  - The input sequence (e.g., a sentence) is fed into the encoder. The encoder processes this sequence through RNNs (or LSTMs, GRUs) and outputs a context vector that summarizes the input sequence.
- Context Vector:
  - The context vector encapsulates the entire input sequence's relevant information. In the traditional Seq2Seq model, this context vector is passed as a fixed-size vector to the decoder.
  - It represents the information needed by the decoder to generate the output.

Seq2Seq: Working

- Decoder:
  - The decoder generates the output sequence token by token, with each step relying on:
    - The context vector (from the encoder).
    - The previous token (or in the case of the first token, a start token or some initial state).
- Output Sequence:
  - The decoder produces an output sequence, which could be of the same length as the input sequence (for tasks like text translation) or different (for tasks like summarization or image captioning).

Seq2Seq: Example of Translation

- Input (English sentence): "How are you?"
- Encoder: The encoder processes the sentence "How are you?" token by token and encodes it into a context vector.
- Context Vector: The context vector stores the important features of the sentence.
- Decoder: The decoder uses the context vector to generate the French translation token by token.
- Output (French sentence): "Comment ça va ?"

Seq2Seq: Advantages

- Handling Variable-Length Input and Output Sequences:
  - Seq2Seq models can handle sequences of varying lengths for both input and output, making them suitable for tasks like translation where sentences may vary in length.
- Flexible Architecture:
  - The encoder-decoder architecture can be adapted with different types of models, such as attention mechanisms or transformers.
- Multilingual and Cross-Domain Applications:
  - Once trained, Seq2Seq models can be used for multiple domains (e.g., translating between different languages) with suitable modifications.

Thank you

This presentation is created using LibreOffice Impress 7.4.1.2, can be used freely as per GNU General Public License

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com