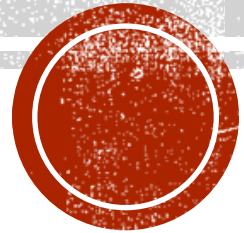
MODULE 5
REGION-BASED CNNS



Dr Divya Meena Sundaram

Sr. Assistant Prof. Grade 2

SCOPE

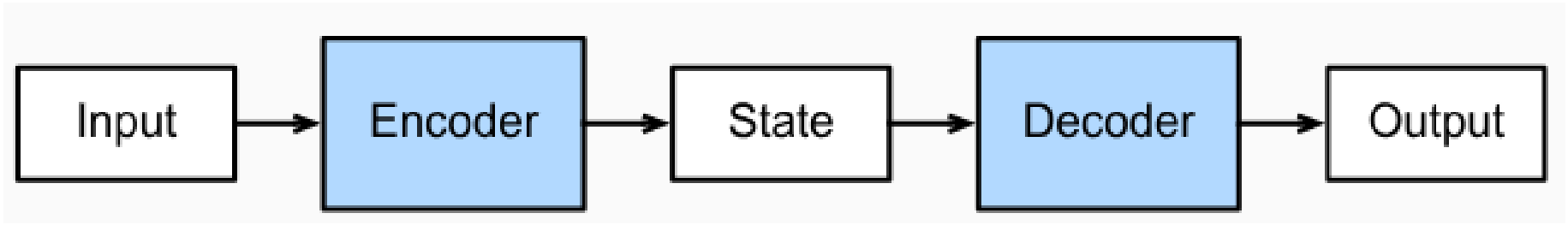
VIT-AP University

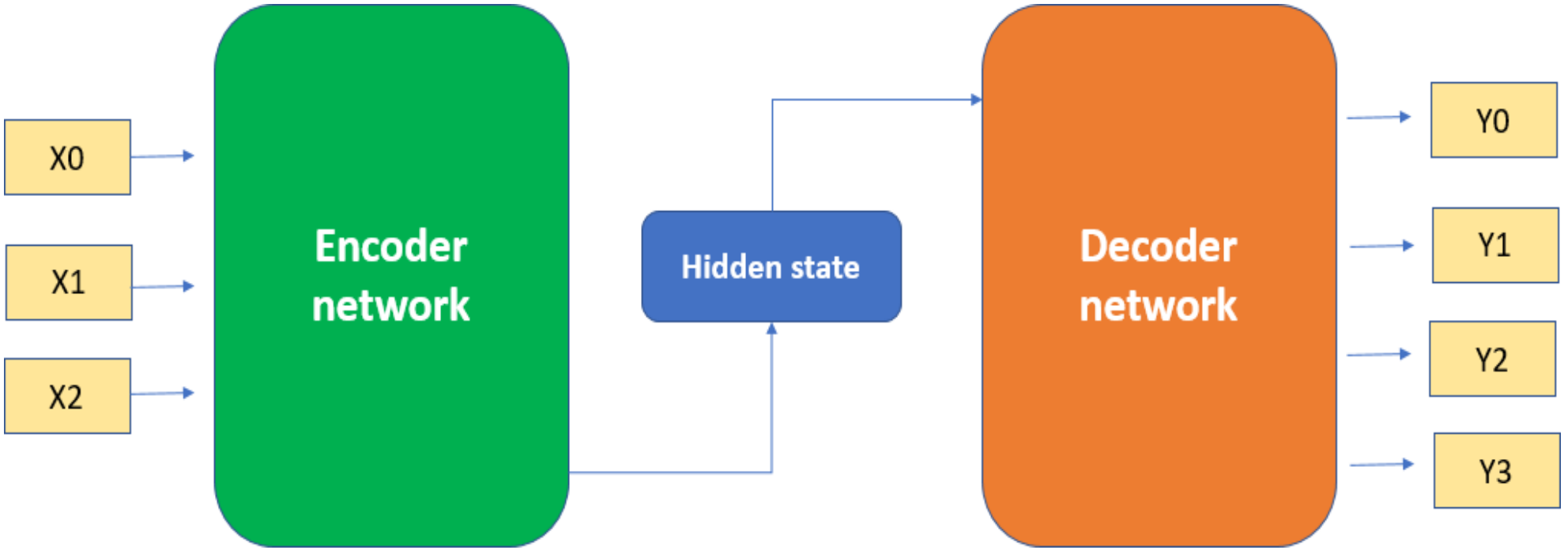
Encoder-Decoder Models, Attention approaches, RCNN, Yolo and its versions-Data Collection, Image labeling and Training. Build Custom models, Comparative analysis. Various Applications

Encoder-Decoder Architecture

The encoder-decoder architecture is a common design used **for tasks where input and output sequences have different lengths** and **are not directly aligned**, such as language translation, text summarization, and chatbots.

1. **Encoder:**
    - The encoder **reads the input sequence** (word by word or token by token).
    - It **summarizes** the input into a fixed-length vector (called a **context vector** or **latent representation**).
2. **Decoder:**
    - The decoder **takes the encoded vector** and generates the output **one word at a time**.
    - It predicts each word **based on the encoded input and previous words** in the output.

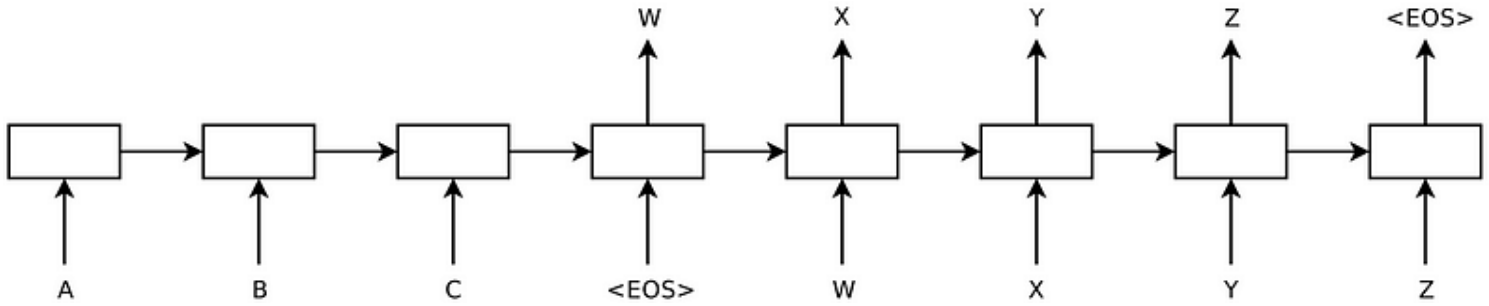

Input
Encoder
State
Decoder
Output


X0
X1
X2
Encoder network
Hidden state
Decoder network
Y0
Y1
Y2
Y3

1. **Encoder (Understanding the Story)**
    - The **encoder reads** the input sentence **word by word** and converts it into numbers (a hidden state).
    - This hidden state is like a **summary** of the whole input.
2. **Hidden State (Memory of the Story)**
    - The hidden state is the **numerical representation** of the input.
    - It contains the **important meaning** of the input sentence.
3. **Decoder (Retelling the Story in Another Language)**
    - The **decoder takes** the hidden state as input.
    - It generates the output **one word at a time**, using both the hidden state and previous words it has generated.

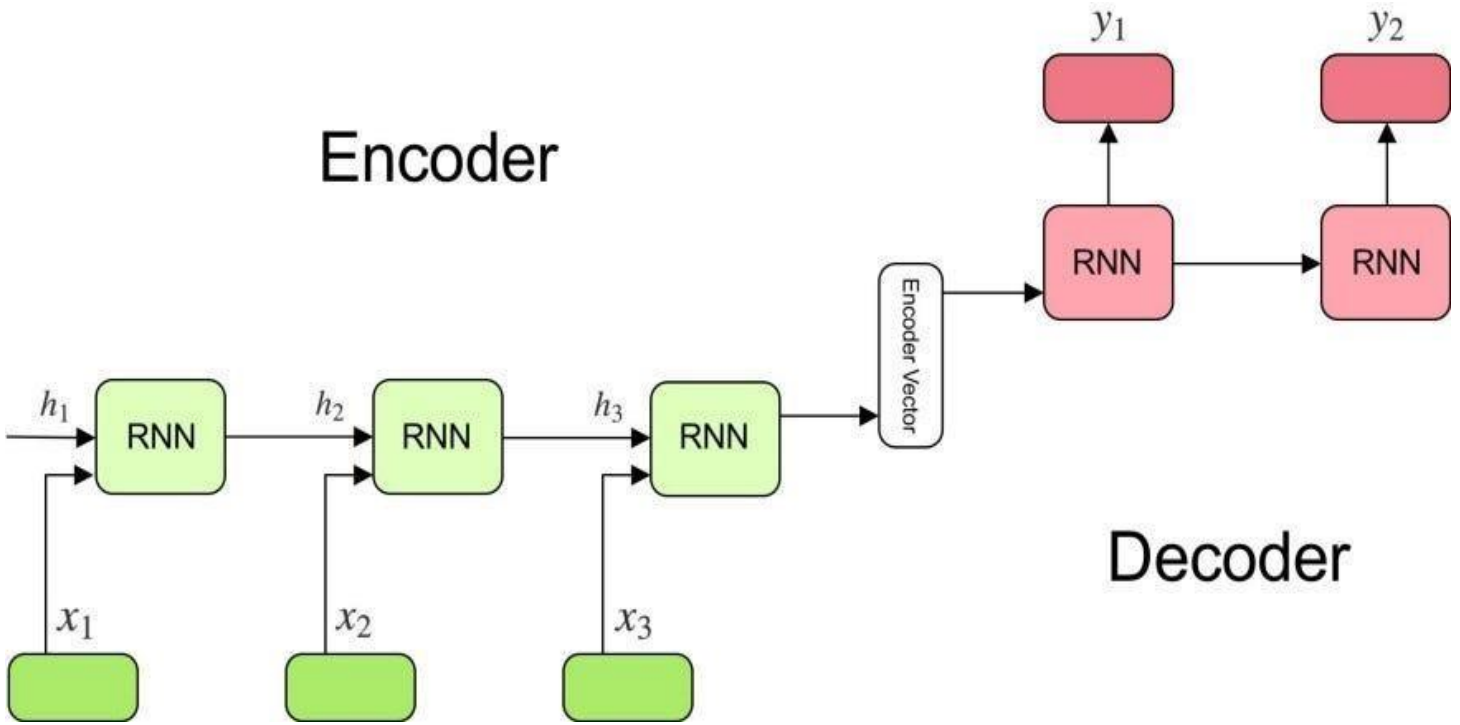
A Basic Approach to the Encoder-Decoder Model

- Initially machine translation (MT) problems were faced using statistical approaches, based mainly on Bayes probabilities.
- But when neural networks became more powerful and popular, researchers began to explore the capabilities of this technology, and new solutions were found. It is called neural machine translation (NMT).



- The encoder, on the left hand, receives sequences from the source language as inputs and produces, as a result, a compact representation of the input sequence, trying to summarize or condense all of its information.
- At each time step, the decoder generates an element of its output sequence based on the input received and its current state, as well as updating its own state for the next time step.

- Let's take machine translation from English to French as an example.
- Given an input sequence in English: "They", "are", "watching", ".",
- This encoder-decoder architecture first encodes the variable-length input into a state, then decodes the state to generate the translated sequence, token by token, as output: "Ils", "regardent".



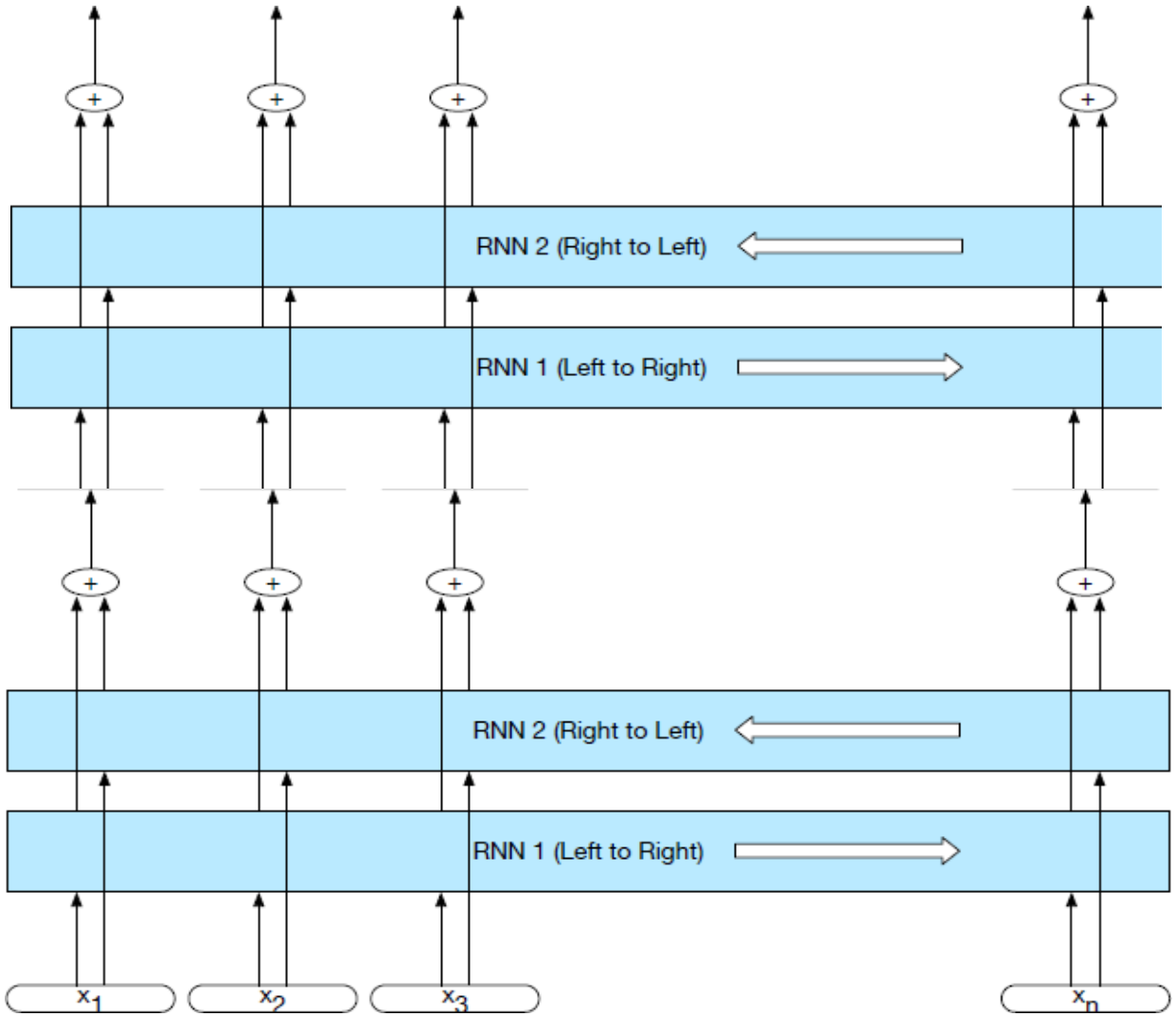
The Architecture of Encoder-Decoder

The Encoder

- The encoder is basically LSTM/GRU cell.
- Layers of recurrent units where, in each time step, an input token is received, collecting relevant information and producing a hidden state.
- In an LSTM, the unit combines the current hidden state and input, processes them through gates, and produces a new hidden state and cell state. The hidden state can be used as output or passed to the next time step.
- Popular architectural choices for Encoder is Stacked Bi-LSTMs.

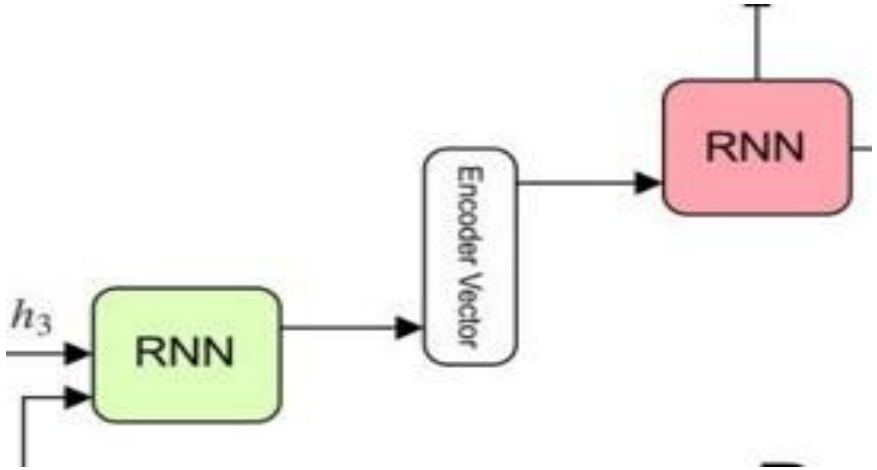
Stacked Bi-lstms

- A widely used encoder design makes use of stacked Bi-LSTMs.
- The first Bi-LSTM layer processes the input sequence and generates hidden states from both the forward and backward passes. These hidden states are then concatenated and passed as input to the next Bi-LSTM layer. This process continues through multiple layers, each refining and enriching the contextual representations at every time step, leading to a deeper understanding of the sequence.



The Encoder Vector (Intermediate Vector)

- The encoder vector is the last hidden state of the encoder, and it tries to contain as much of useful input information as possible to help the decoder get the best results.
- It's the only information from the input that the decoder will get.

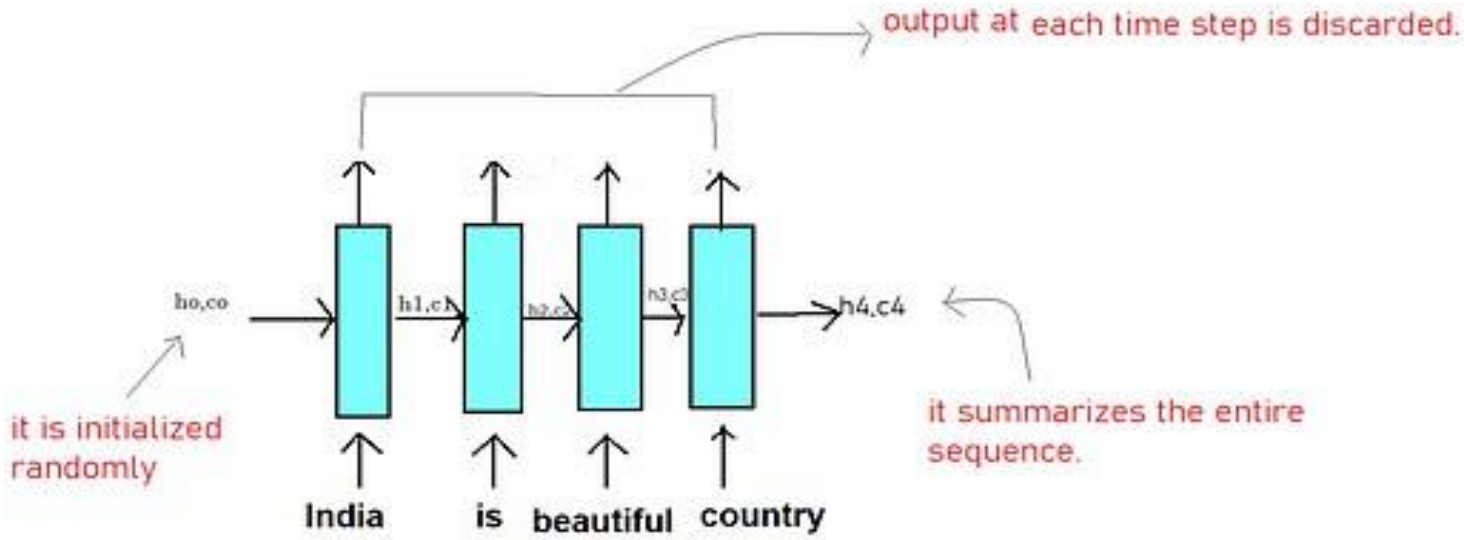


The Decoder

- Layers of recurrent units — e.g., LSTMs — where each unit produces an output at a time step t.
- The hidden state of the first unit is the encoder vector, and the rest of the units accept the hidden state from the previous unit.
- The output is calculated using a softmax function to obtain a probability for every token in the output vocabulary.
- Unlike encoders, decoders unfold a vector representing the sequence state and return something meaningful for us like text, tags, or labels.

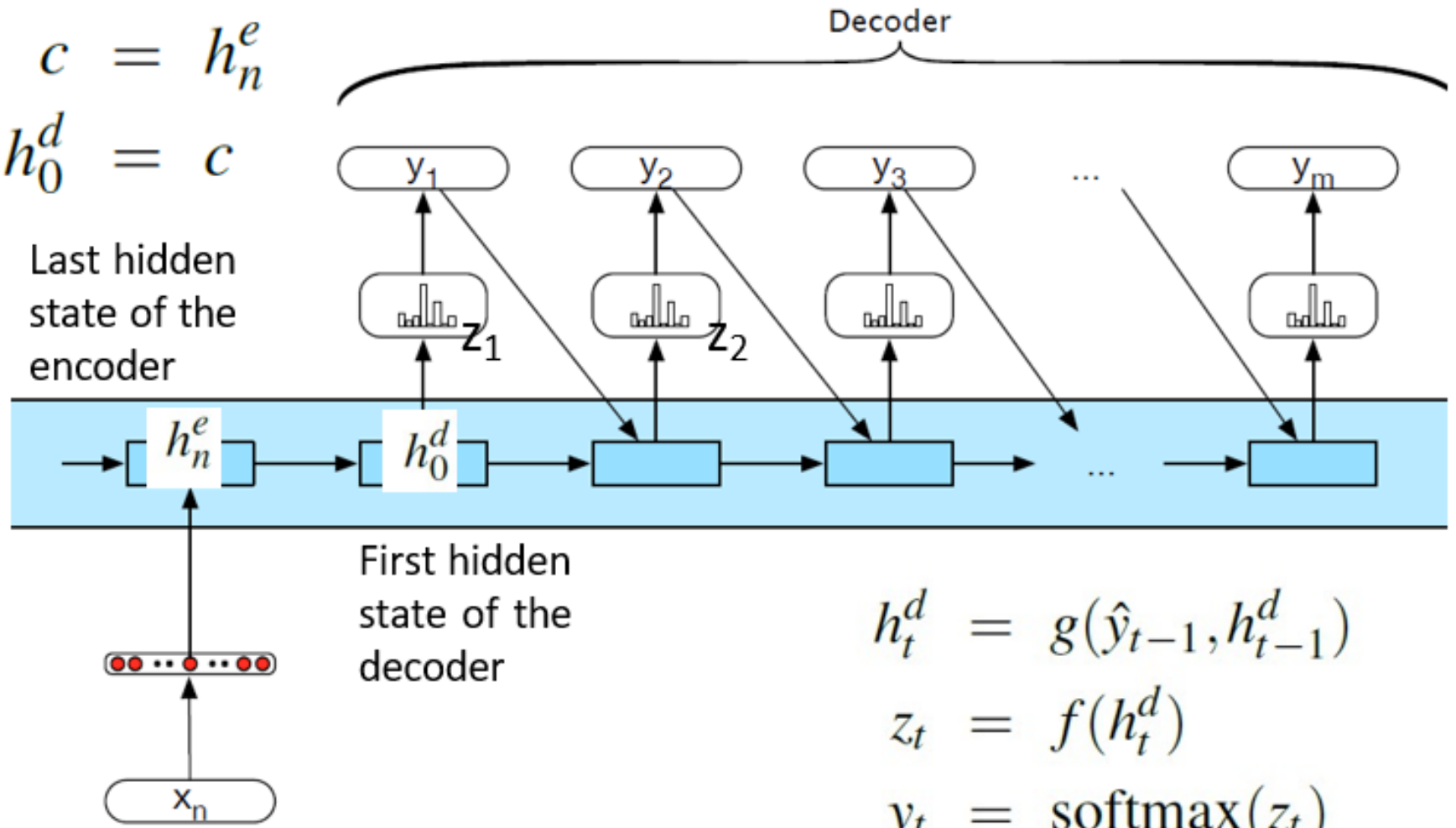
The Decoder

- An essential distinction with encoders is that decoders require both, the hidden state and the output from the previous state.
- When the decoder starts processing, there's no previous output, so we use a special token <start> for those cases.
- The decoder continues generating tokens until it produces an end-of-sequence marker (EOS), indicating the sequence is complete.



Decoder Basic Design

- Autoregressive generation is used to produce an output sequence, an element at a time, until an end-of-sequence marker is generated.



Limitations of Encoder-Decoder Architecture

- The final numerical representation or hidden state in the encoder network has to represent the entire context and meaning of a sequence of data.
- If the sequence of data is long enough, it may get challenging and the information about the start of the sequence might get lost in the process of compressing the entire information in the form of numerical representation.

Applications of Encoder Decoder

- Make-a-Video: Recently introduced AI system by Facebook / Meta namely Make-a-Video is likely powered by deep learning techniques, possibly including an encoder-decoder architecture for translating text prompts into video content.
- Machine translation
- Image captioning
- Speech Recognition
- Text Summarization

Summary

- The encoder-decoder architecture has become a popular and effective tool in deep learning, particularly in the fields of natural language processing (NLP), image processing, and speech recognition.
- Encoder-decoder architecture can be combined with different types of neural networks such as CNN, RNN, LSTM, etc. to enhance its capabilities and address complex problems.