FLOW

1. What is ASR?
2. Motivation for ASR
3. Paper Overview in short
4. Dataset
5. ASR Block Diagram
6. Technical depth of RNN and LSTM
7. Timeline (workplan) of work for stage-2

What is ASR?

1. Speech recognition is inter-disciplinary sub-field of computational linguistics.
2. It develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers.
3. It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Motivation (Why ASR?)

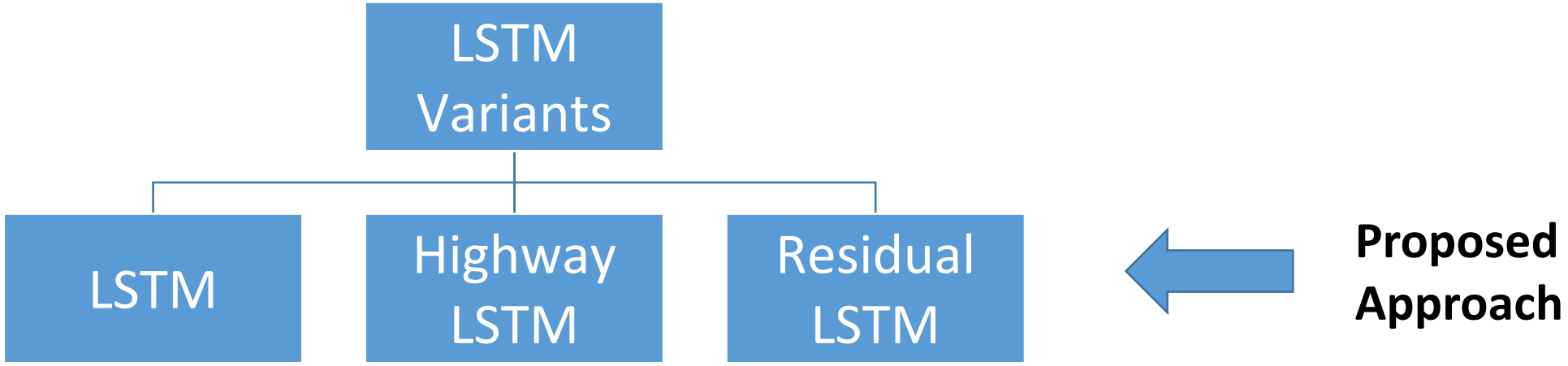
- **ASR has many applications such as -**

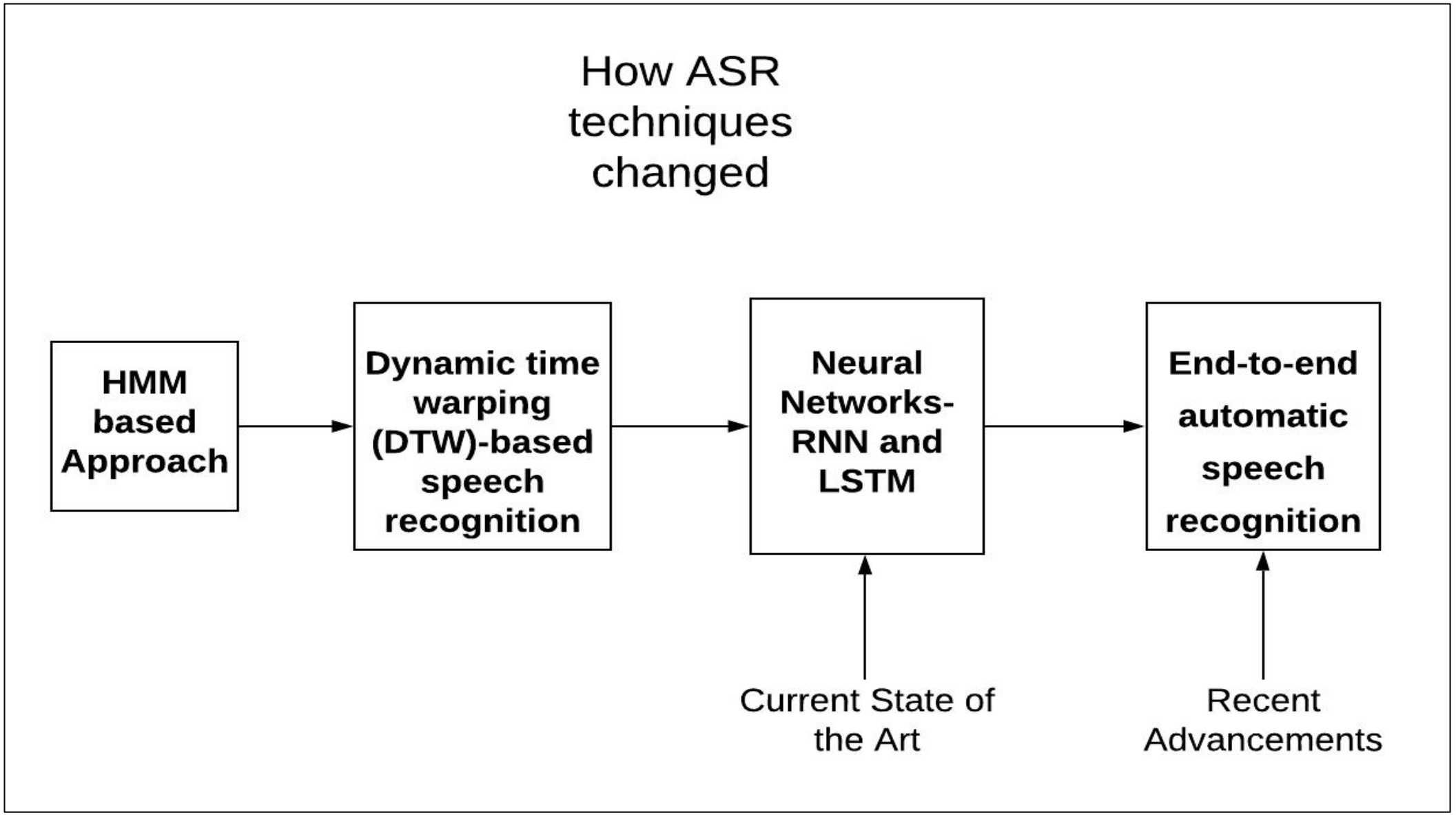
1. In-car systems (Voice Commands)
2. Health care (Medical documentation)
3. Military (Fighter Aircraft(Voice Commands))
4. Usage in education and daily life (Ex. Blind students can benefit)
5. Hands-free computing (Ex. Coding by just voice commands)
6. Home automation and many more applications exist.

Overview:

In this paper, they proposed a novel architecture for a deep RNN:

RESIDUAL LSTM(Variant of Long Short Term Memory)




How ASR techniques changed
HMM based Approach
Dynamic time warping (DTW)-based speech recognition
Neural Networks-RNN and LSTM
End-to-end automatic speech recognition
Current State of the Art
Recent Advancements

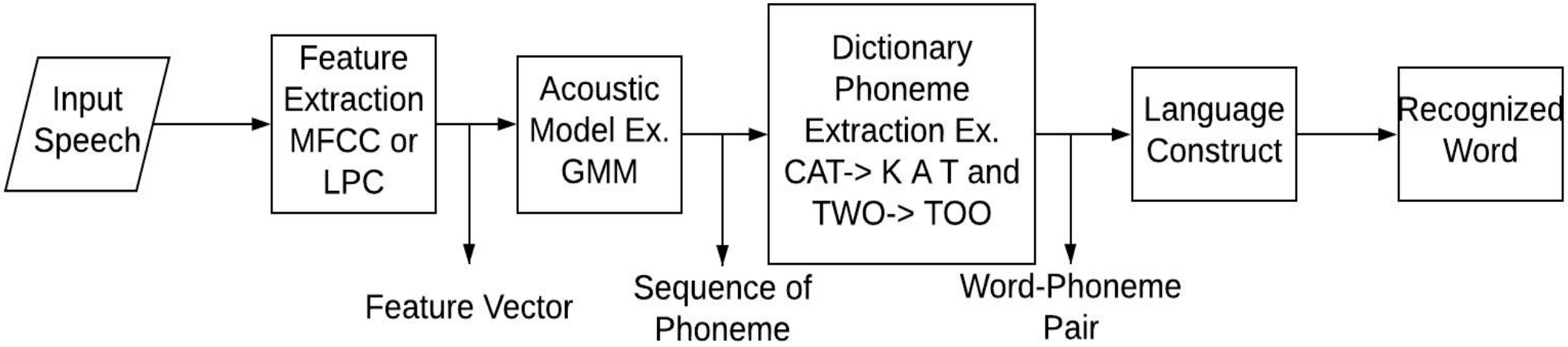
Dataset- AMI Meeting Corpus

- The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings.
- The meetings were recorded in English using three different rooms with different acoustic properties.
- Example Audio from dataset-

Figure 1: AMI's three instrumented meeting rooms.

Source: http://groups.inf.ed.ac.uk/ami/corpus/

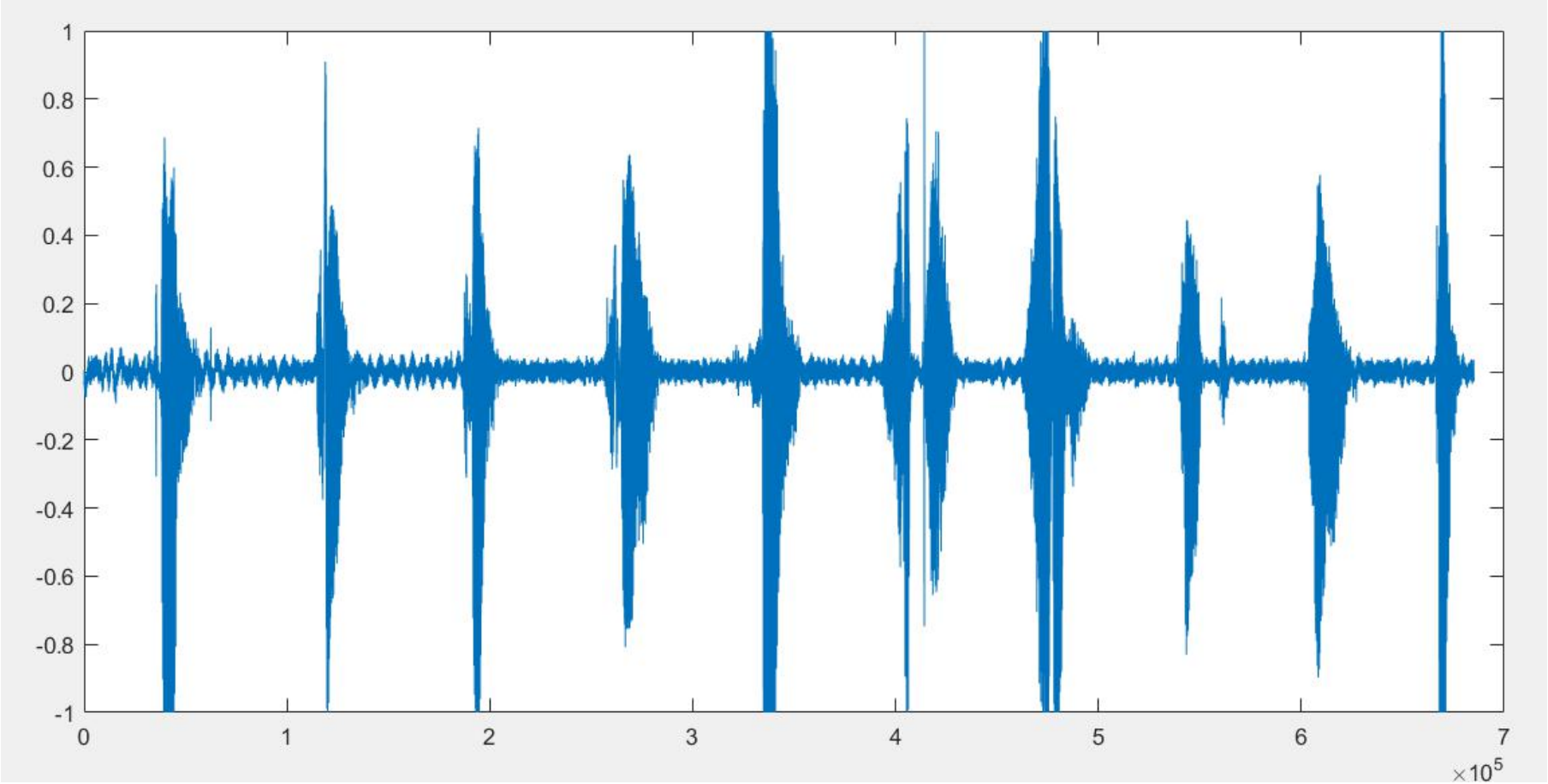
Abstract View- ASR



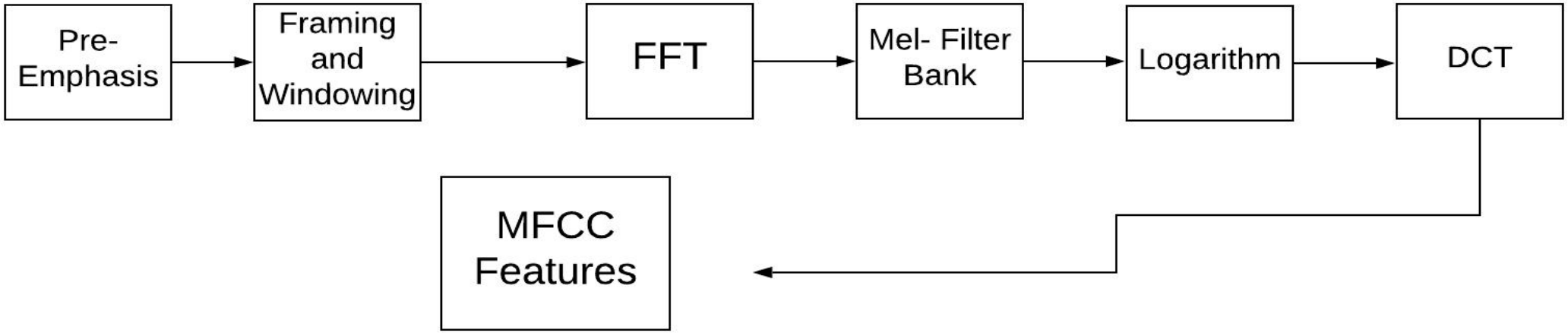
Block Diagram of Speech Recognition

Reference: https://www.cse.iitb.ac.in/~nirav06/i/HMM_Report.pdf

Plot of Sampled Data -Y of audio sample



MFCC FEATURES



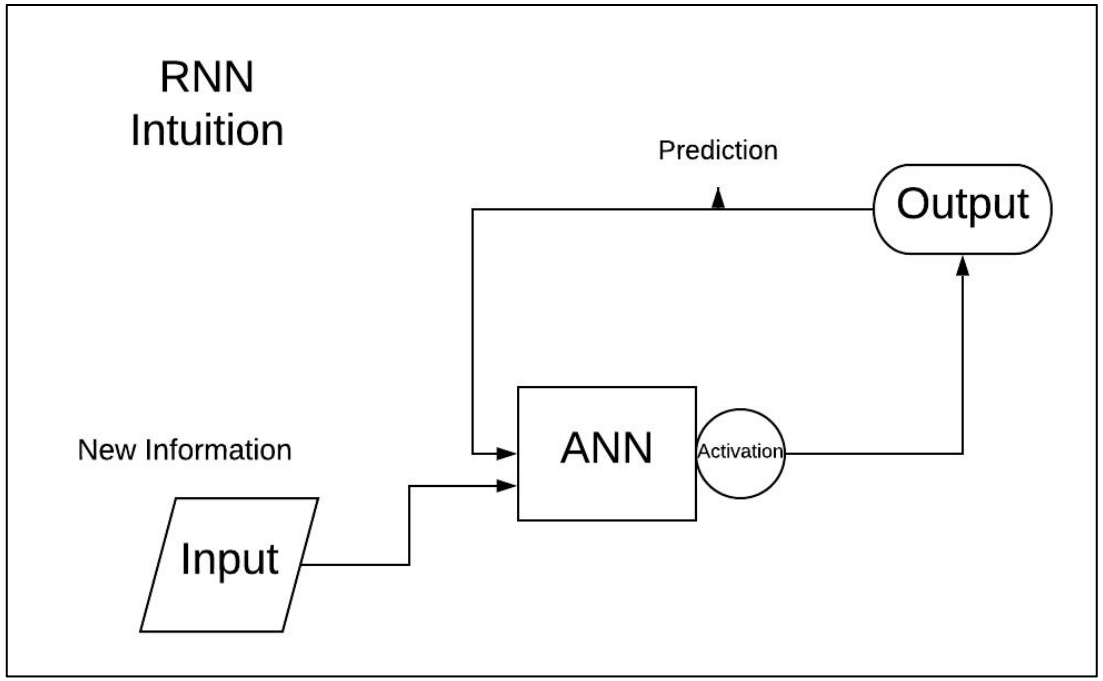
Block Diagram of MFCC

1. Important Content of speech is Linguistic content.
2. Unimportant- Background noise, emotions, silence etc.

Reference -
Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

RNN- Recurrent Neural Network

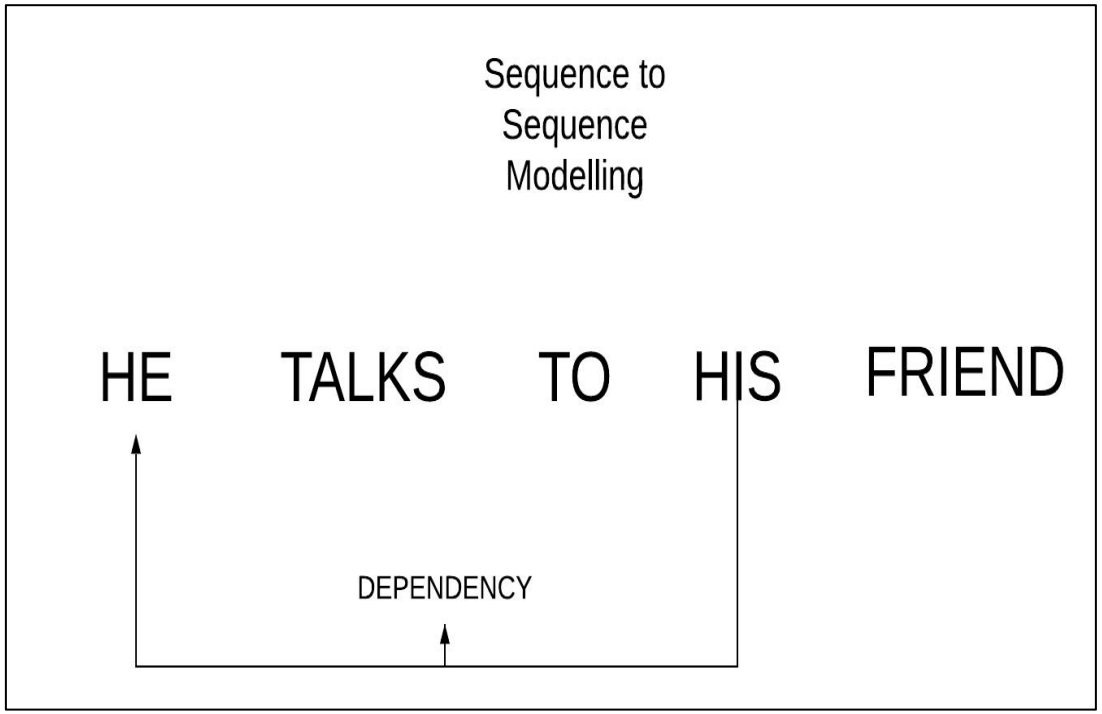
A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to **process sequences of inputs**.



Why RNN? or LSTM (a variant of RNN)

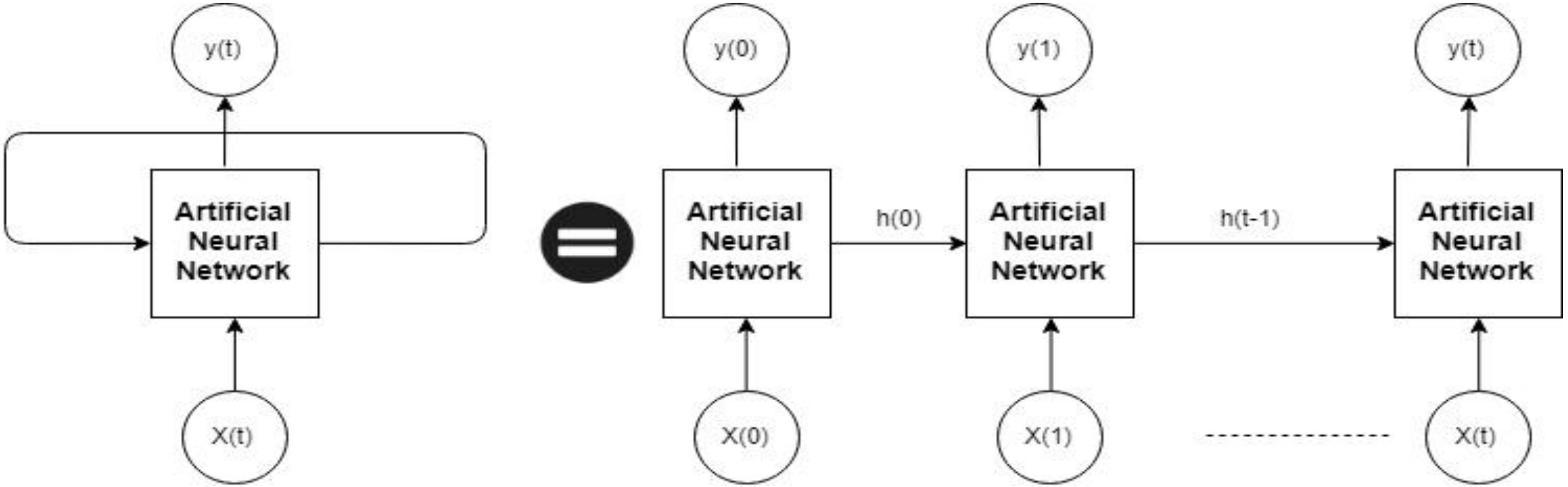
- The past input data influences the current output- Example

1. **Stock Price Prediction-** Output depends on previous data.
2. **Text Generation**- As shown in right.
3. **Automatic Speech Recognition**



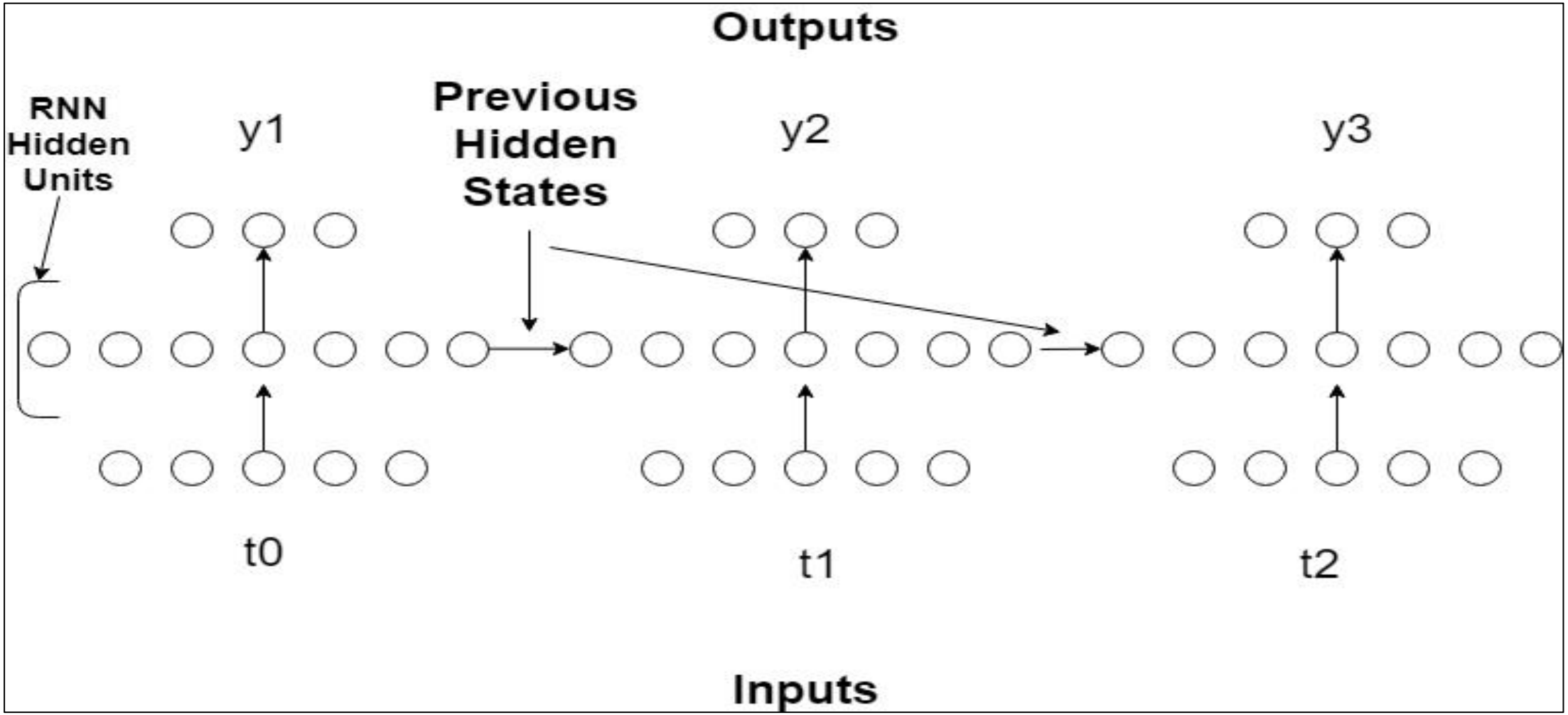
About RNN

1. RNN is a multiple copy of the same network (For ex. ANN), that receives inputs at different times as well as it's previous hidden state
2. This diagram shows it's unrolling.



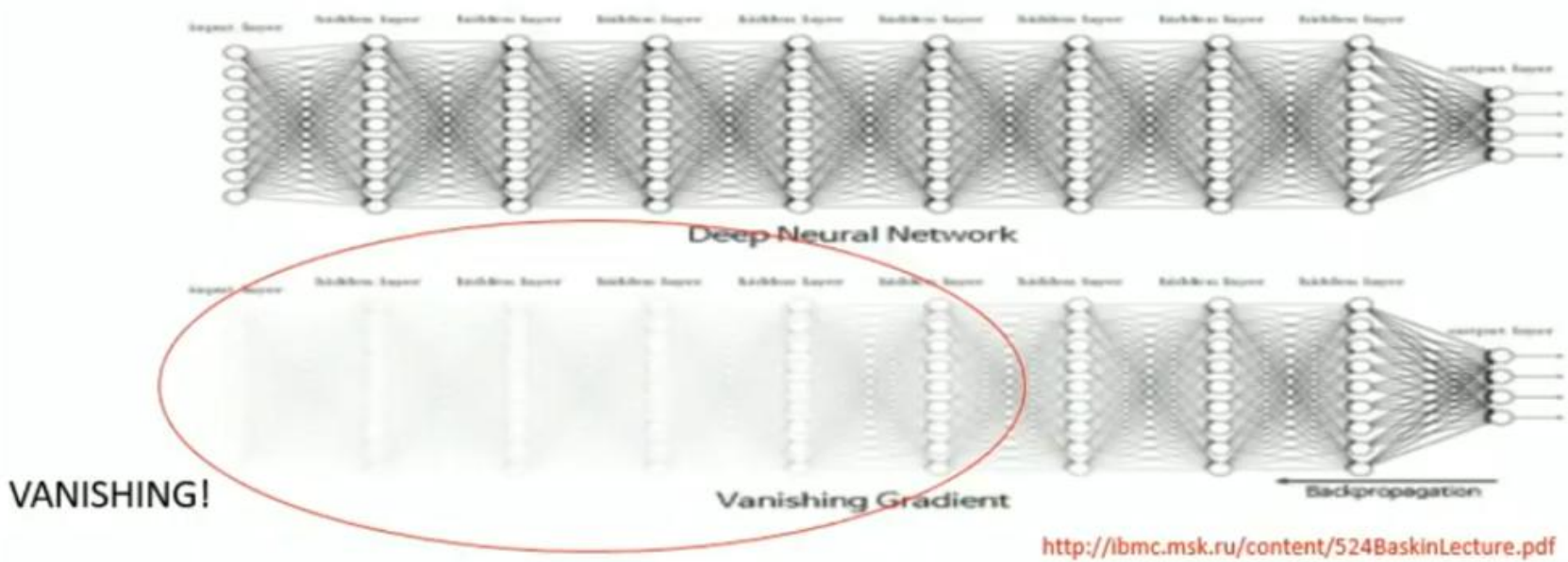
Ex. CAT phonemes are K,A,T. It needs 3 inputs and hence three ANN units in RNN.

More about RNN



Disdadvantage of Feedforward Networks and RNN

Vanishing Gradient Problem



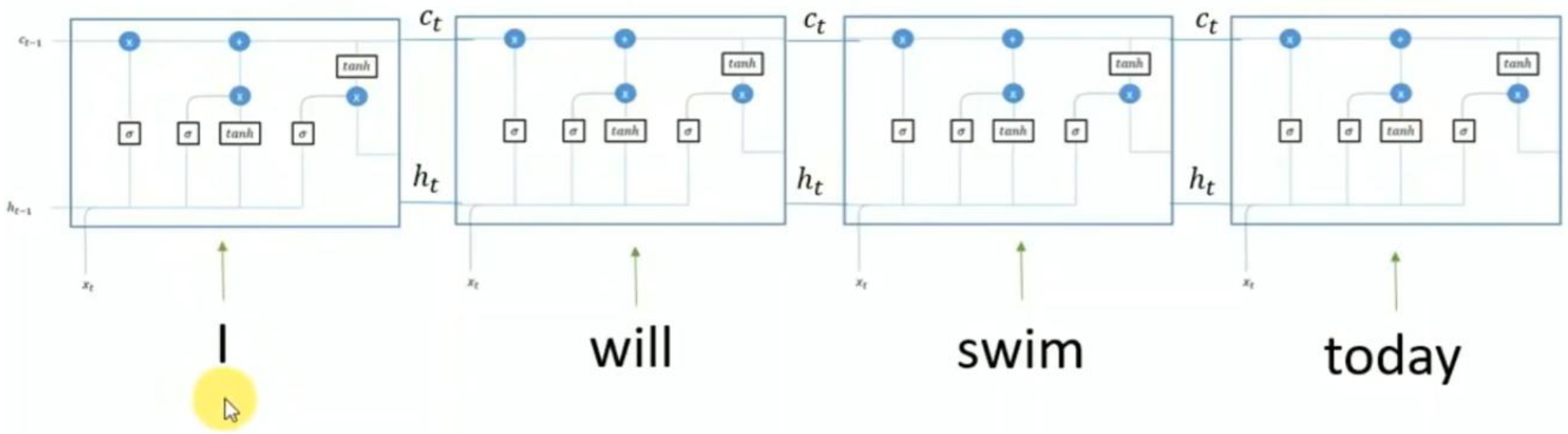
Source: freeCodeCamp.org

Example

Suppose we want to predict:

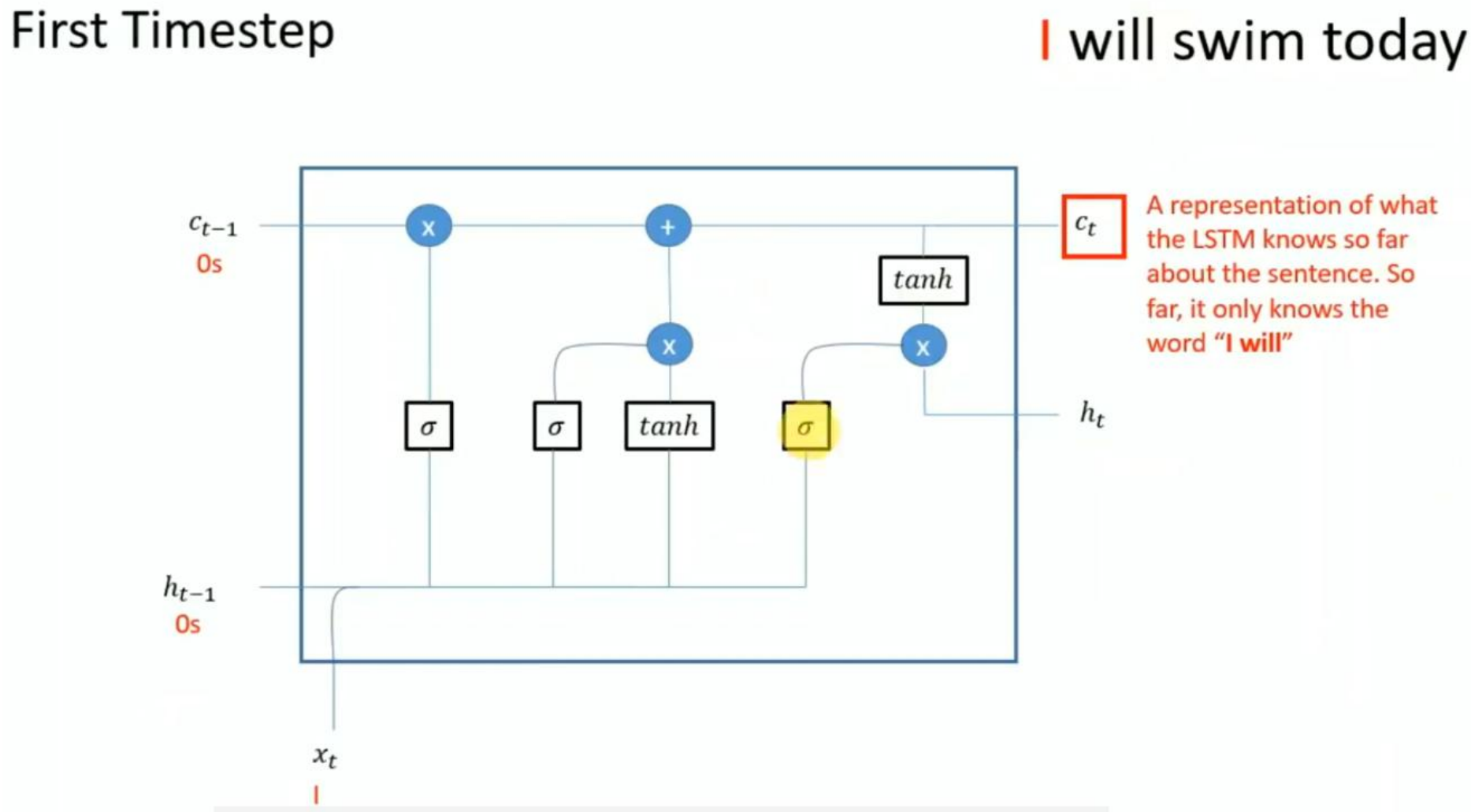
I will swim today

We will need 4 timesteps, since the sentence is composed of 4 words.



Source: freeCodeCamp.org

Ct (Memory cell/Current Output/Context Vector) Description

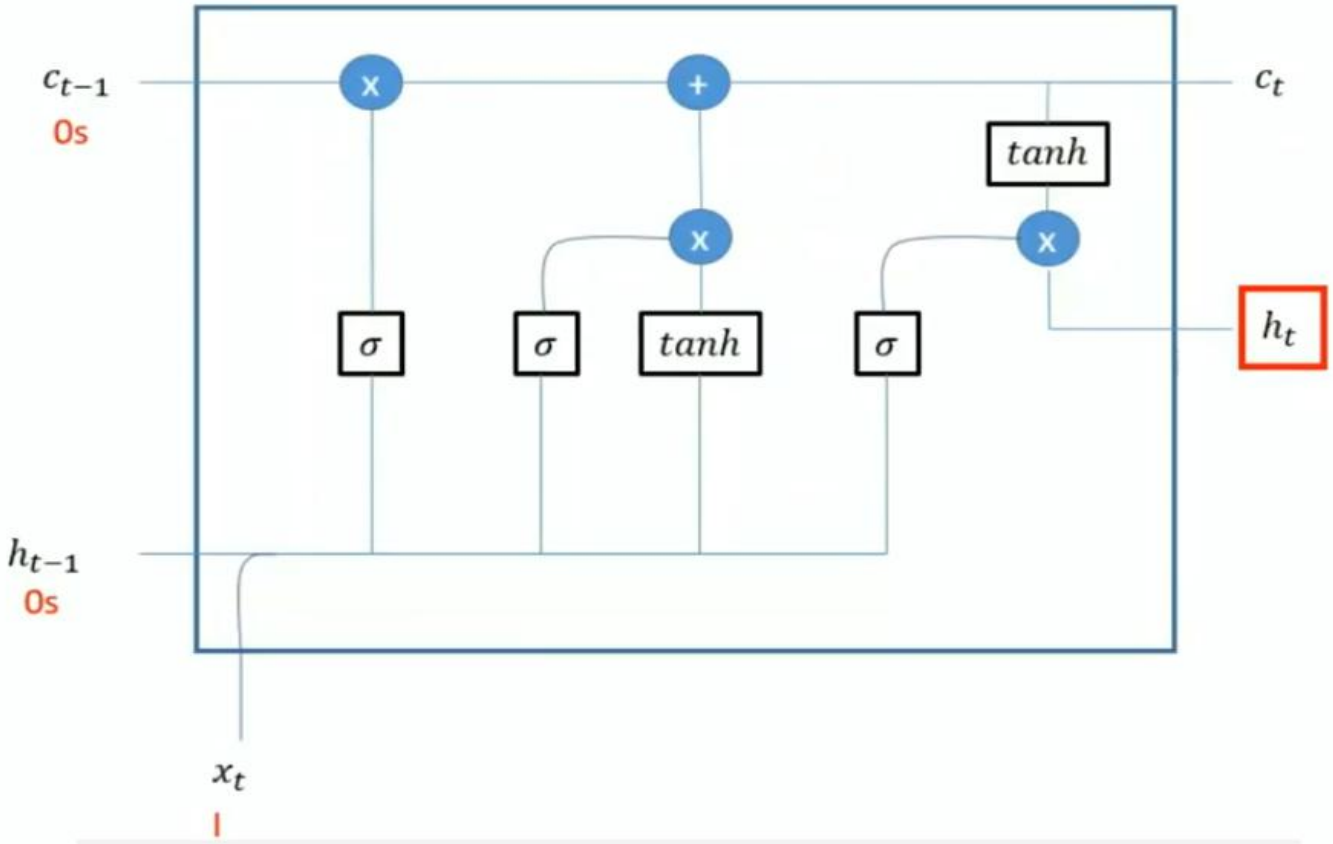


Source: freeCodeCamp.org

ht (hidden output)

First Timestep

I will swim today



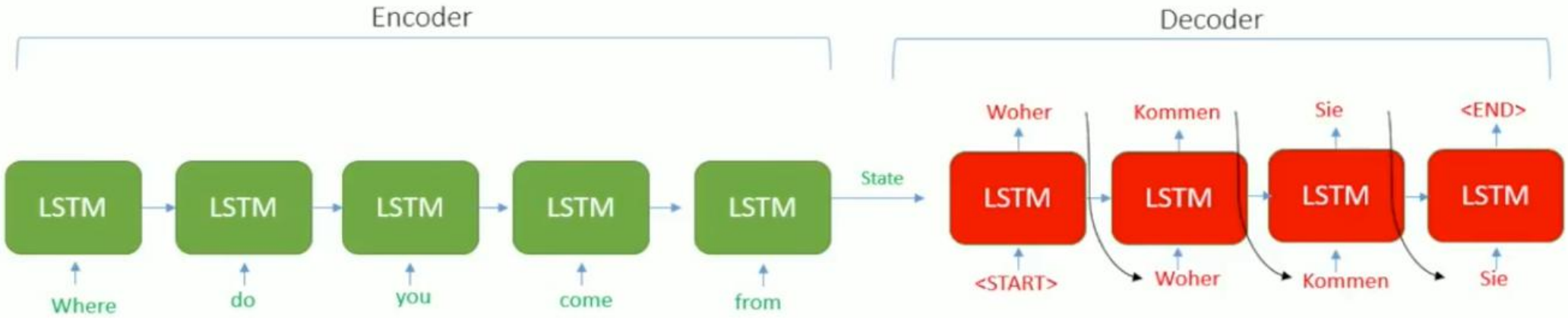
A representation of the output. For example, it may be a representation of the word "***will***"

Source: freeCodeCamp.org

Example - English to German Translation

Seq2Seq

- Machine Translation



Note: The Above Diagram happens at Testing Time

During Training:

Inputs:	**<START>**	WOHER	KOMMEN	SIE
Targets:	WOHER	KOMMEN	SIE	**<END>**

Source: freeCodeCamp.org

Sigmoid VS tanh

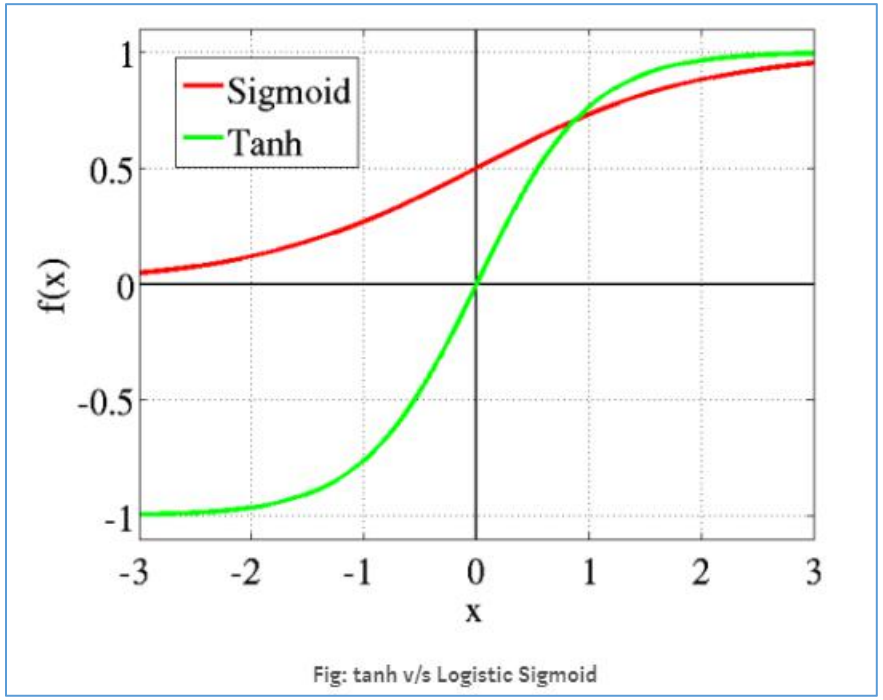


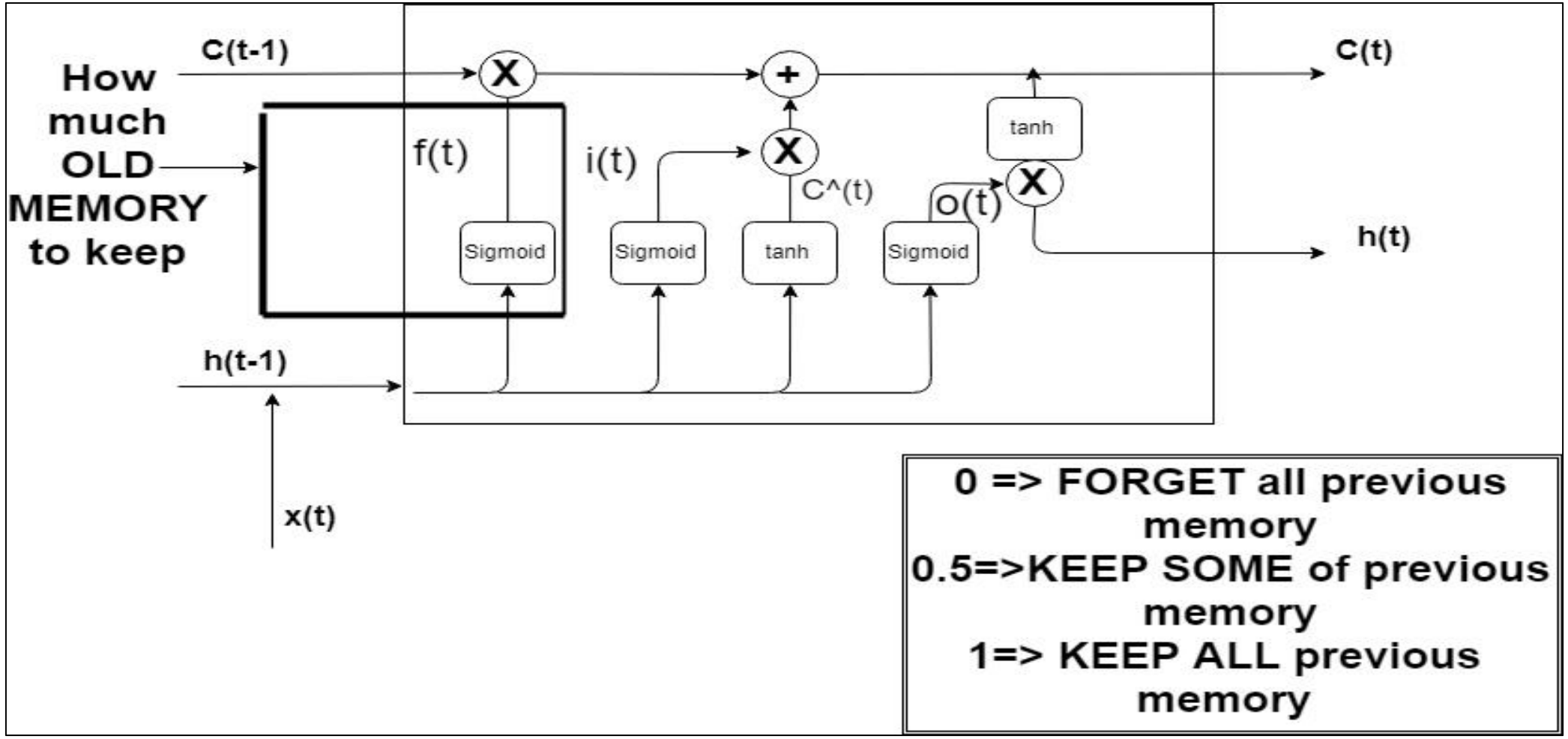
Fig: tanh v/s Logistic Sigmoid

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}.$$

$$\cosh(x) = \frac{1}{2}(e^{x}+e^{-x});\ \sinh(x) = \frac{1}{2}(e^{x}-e^{-x});\ \tanh(x) = \frac{\sinh(x)}{\cosh(x)}$$

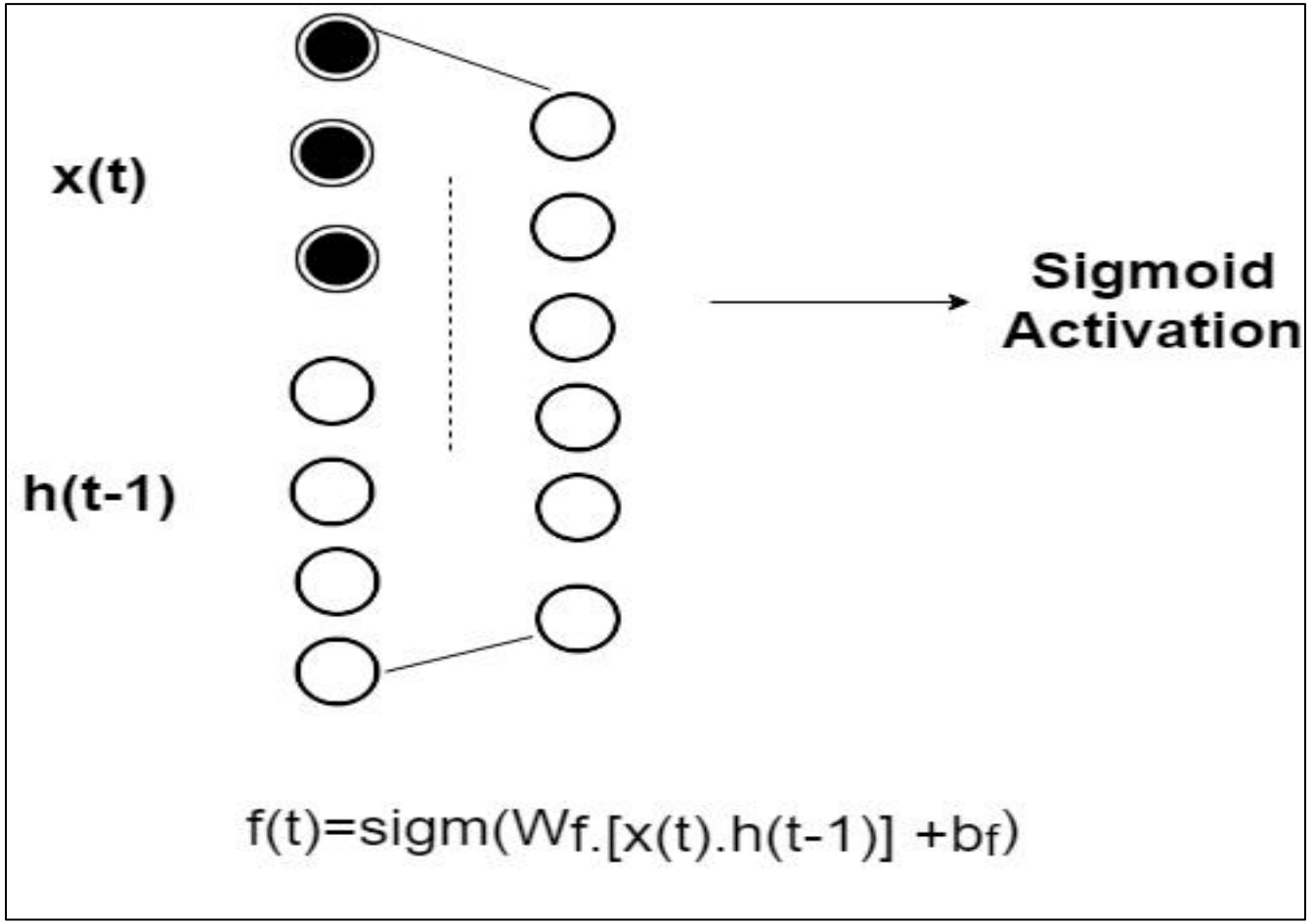
Source: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

LSTM Architecture(Forget gate)

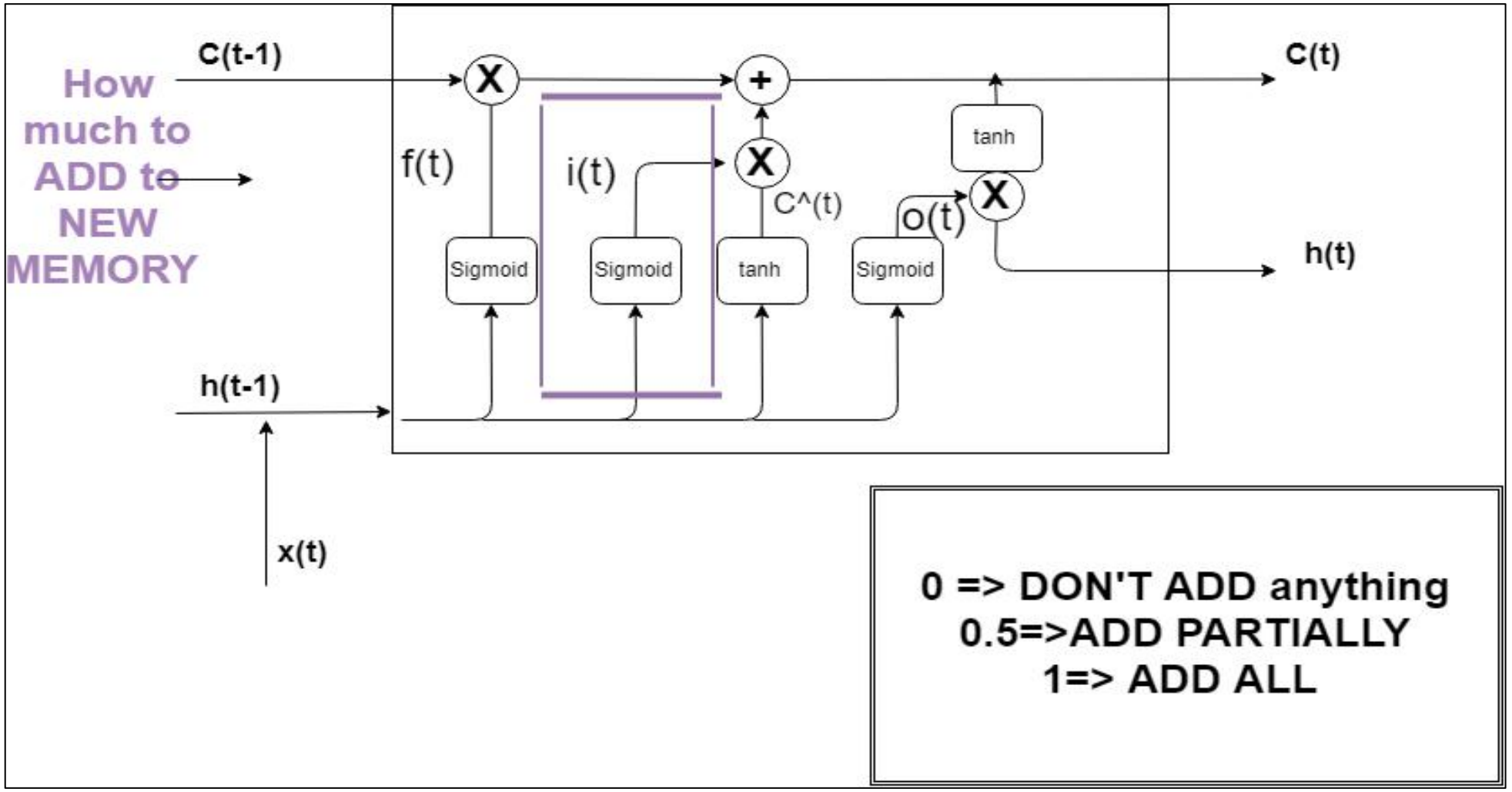


Reference: freeCodeCamp.org

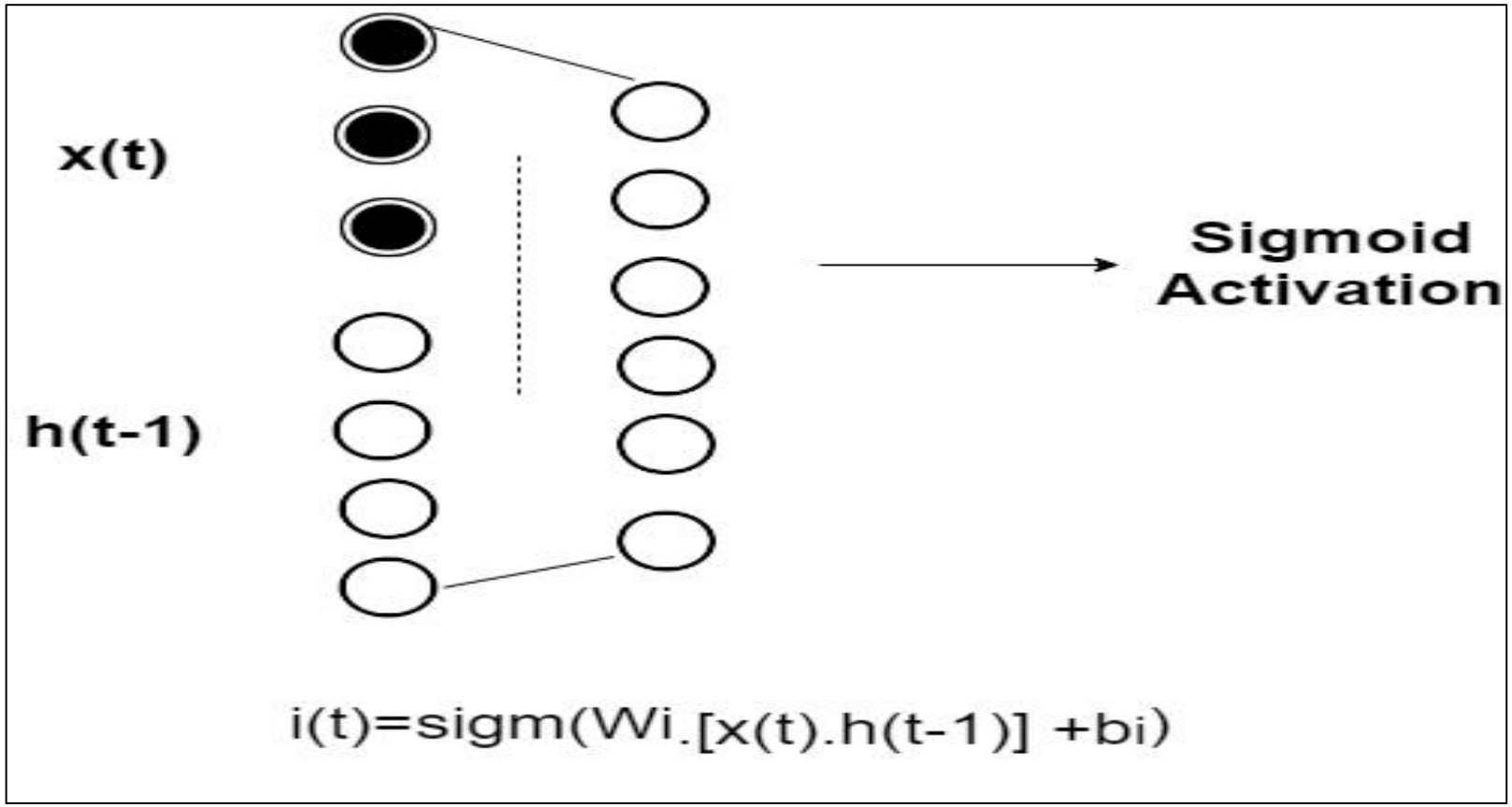
Forget Gate Equation:



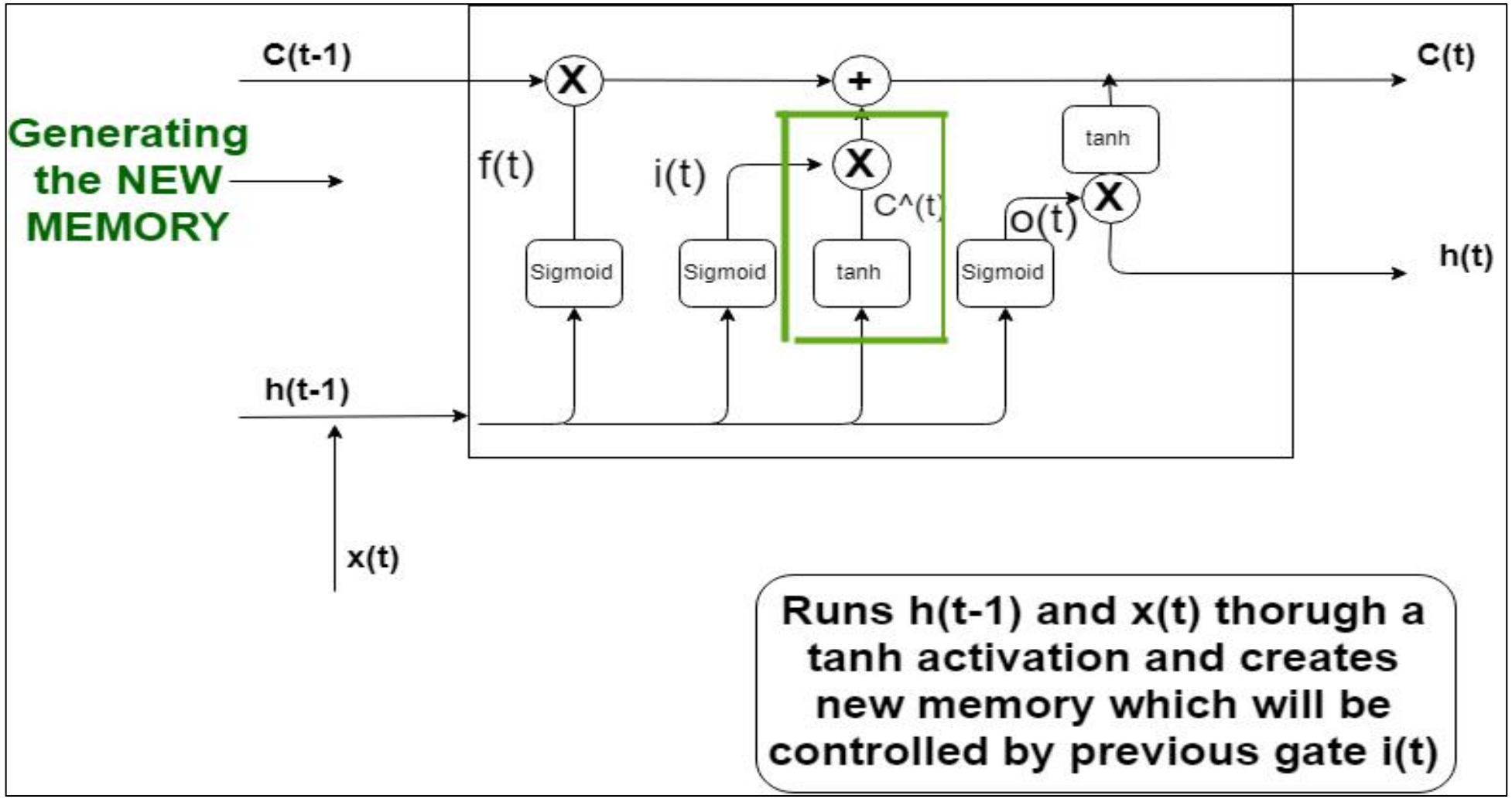
LSTM Architecture(Input gate)



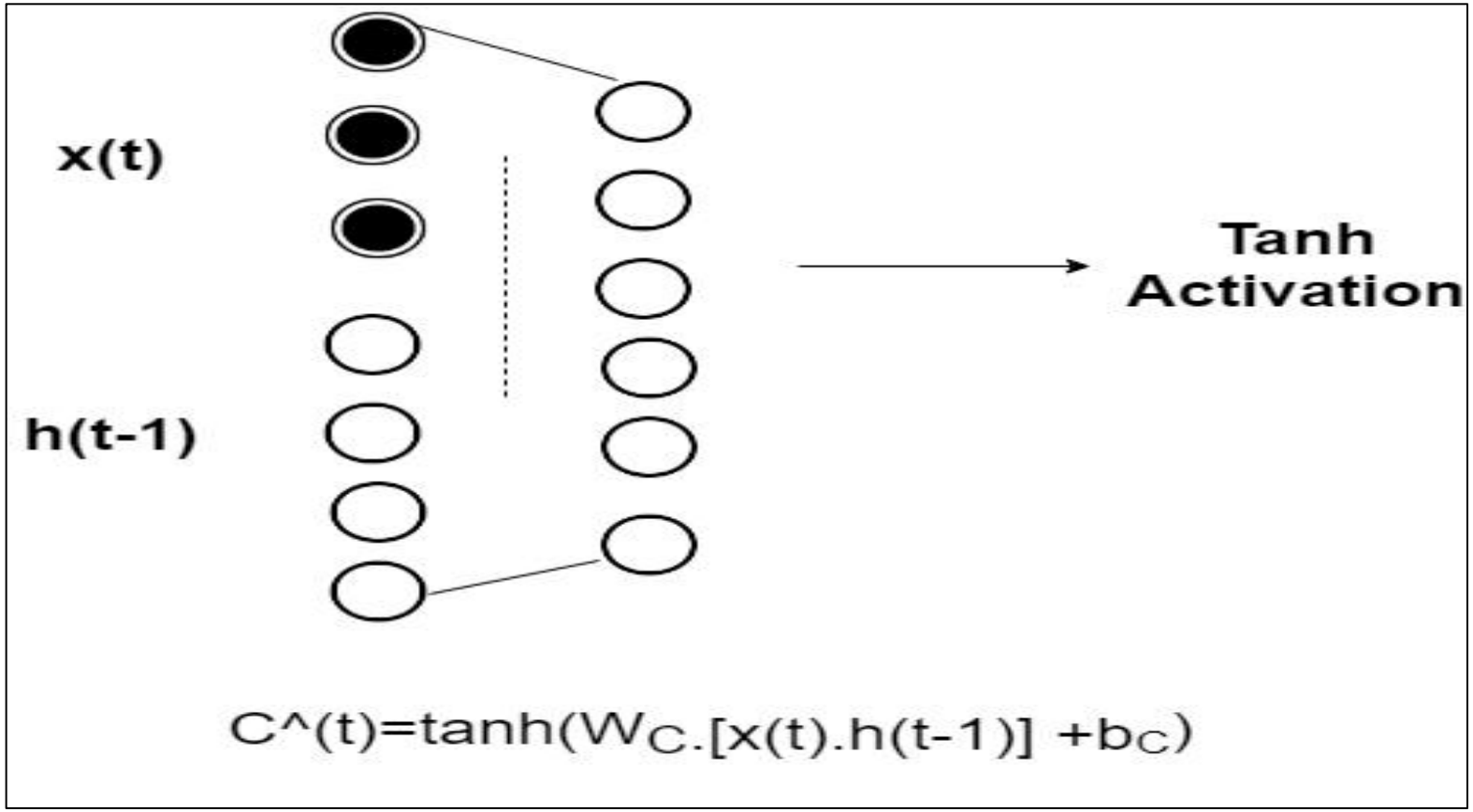
Input Gate Equation



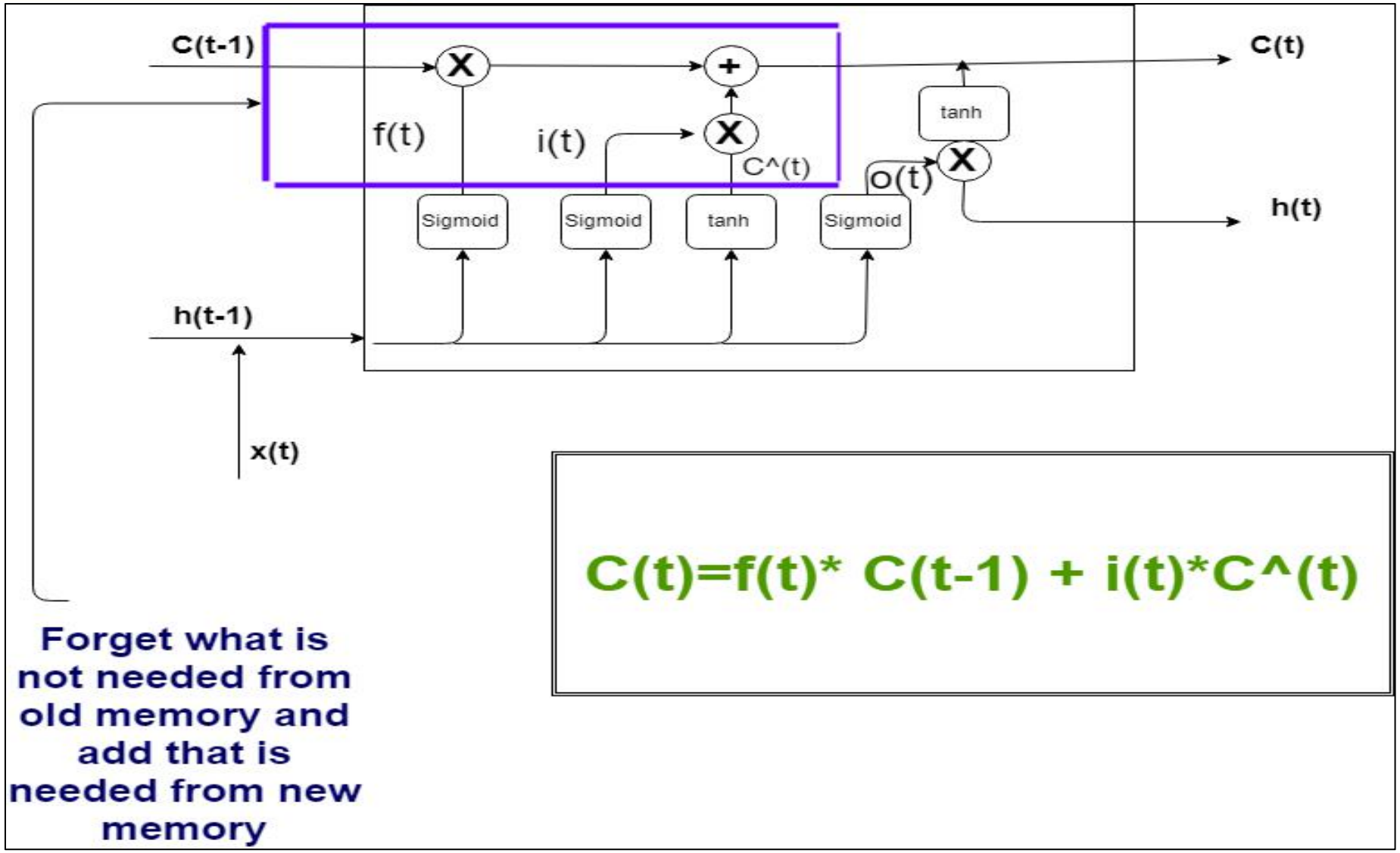
LSTM Architecture(New Memory based on previous hidden state and current input)



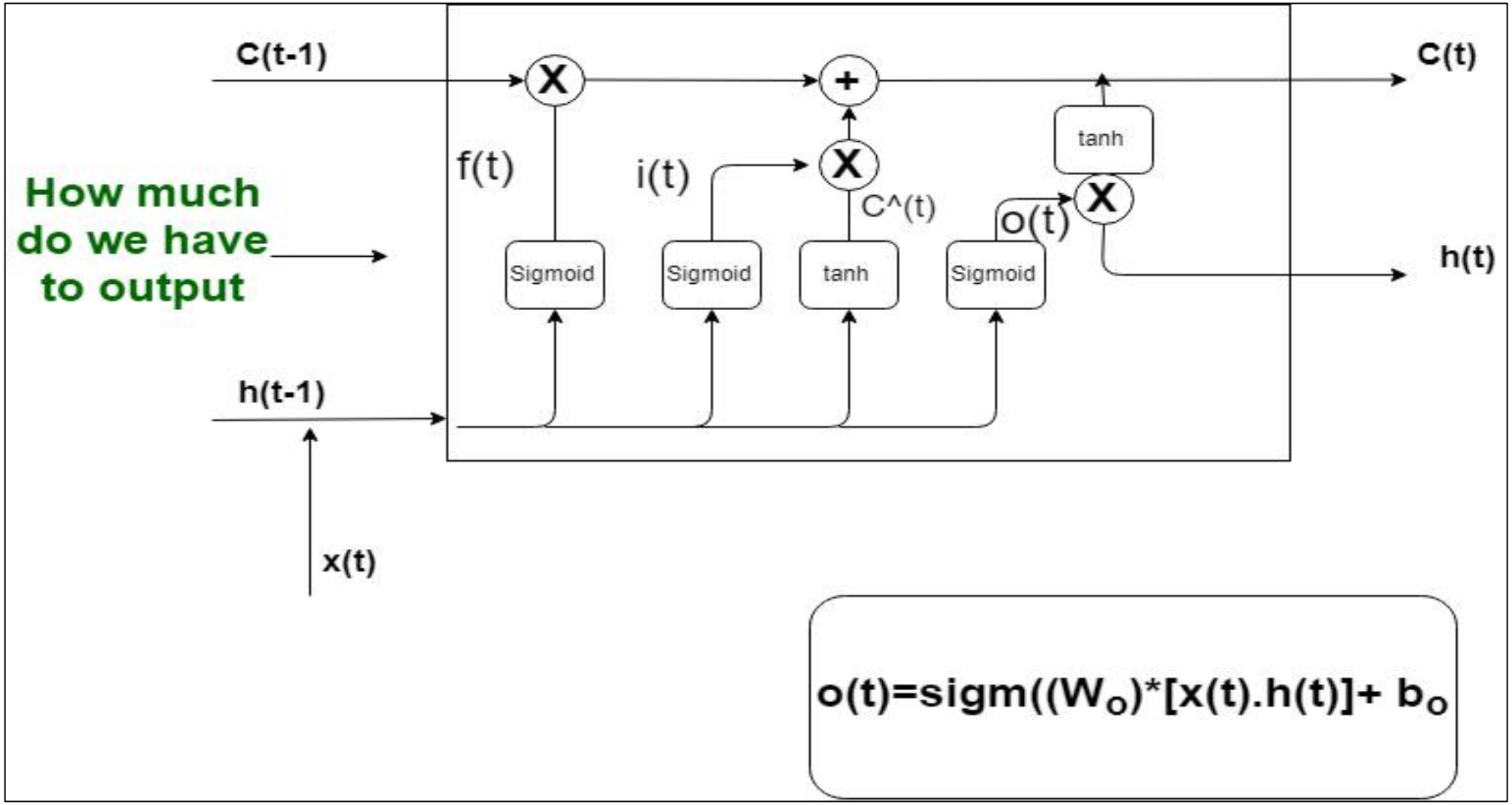
Equation of Current State C^(t)



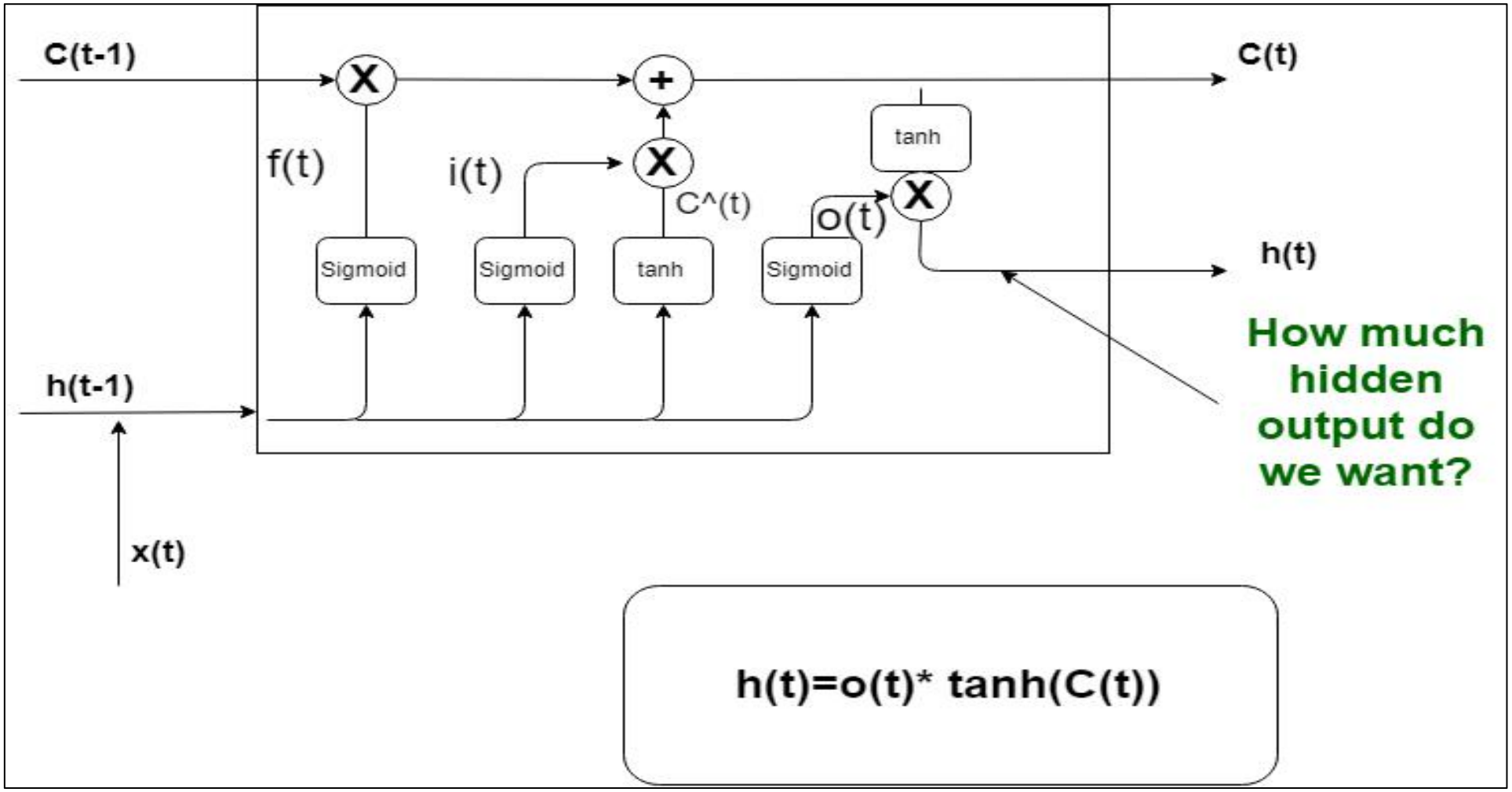
LSTM Architecture(New Memory = previous + current memory)



LSTM Architecture(How much of new memory to be as output)



LSTM Architecture(Hidden State Output)



Thanks