Review of Deep Learning Models from Convolution Neural Networks to Transformers

Sandeep Maan[1], Gian Devi[2]

[1] *Associate Professor of Computer Science, Govt. College for Girls, Sector-14, Gurugram, Haryana-India*

[2] *Assistant Professor of Computer Science, Govt. College for Girls, Sector-14, Gurugram, Haryana-India*

Abstract*— Generative Artificial Intelligence has become synonym for Artificial Intelligence. Specifically, success of Large Language Model (LLM) is going to make long lasting disruptive effect. Application like ChatGPT [1] by OpenAI, BARD by Google are believed to change the human-machine relation in coming years. All these can be attributed to the developments in the field of deep learning during last decade. Things started with Convolutional Neural Networks (CNN) and advancement of GPUs that has made as lasting effect in the field of image processing. In this paper authors have reviewed features and limitation of most three most popular deep learning models viz. Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN) and Transformer model. A systemic study of factors leading to the development of large language models is also presented.

***Index Terms*— Artificial Intelligence, Machine Leaning, Deep Learning, Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN),* Attention, Transformers, Large Language Models (LLM)**

I. INTRODUCTION

The whole field of deep learning is inspired by our neurological system. Our neurological system comprises of basic processing units called neurons. Neurons are connected through axons and dendrites. Synapses acts as interface between two neuron communication [2].

The process happening in Biological (or Natural) Neural Networks was mimicked in Artificial Neural Networks (ANN). ANN forms the basic unit of all deep learning algorithms in one or other way. ANN comprises a number of layers. With each layer itself comprising number of Neurons acting as basic processing unit. First layer of an ANN acts as Input layer to receives input while the last layer act as output layer. The layers in between are termed as hidden layers. Perceptron is the simplest ANN comprising one layer only. While in a Multi Layered Perceptron (MLP) we have multiple layers.

Although ANNs were first used by Marvin Minsky in 1951 and Perceptron was first proposed by Frank Rosenblatt in 1960. Yet for more than half a century they did not find their due place in the field of Artificial Intelligence. This could be attributed to their humongous processing requirements. Moreover, experimental ANN models during 20th century comprised very few hidden layers.

During second decade of 21st century researcher found a way to implement multilayered neural network through GPUs. As GPUs have become powerful by now. It led to advancement of deep learning at a tremendous pace. Within a decade, we have started comparing artificial intelligence with human intelligence. This is directly related to the advancement of Depp Learning. “Deep” in deep learning signifies many layers that comprise the corresponding model.

During this work we compare three most popular deep learning models viz. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Transformer model. We would investigate how limitations of earlier models led to the development of subsequent models. These models are more than responsible in expanding the base of AI from millions (peoples aware of AI) to billions. Now a days everyone is talking about Artificial and machine learning. Hence this paper.

II. CONVOLUTION NEURAL NETWORKS (CNN)

Implementation of CNN architecture(s) coupled with GPU in the field of deep learning have revolutionized the field. The success of Deep Learning as on today can be largely contributed to them. They are broadly used for Image recognition and classification. The majority of successful image processing architectures employ them. They try to mimic cat’s visual cortex [3]. Input to CNN model is a three-dimensional dataset (provided images are not greyscale). Where depth represents amount of parallelism. If input are RGB images then initial depth of model would be 3. Central to CNN are three basic operations applied to input matrices at different depths:

A. Convolution: Convolution operation is central to the CNN model. During this different filter are applied to detect specific features like edge, line, color inside the image.
B. Striding: It represents the number of pixels input matrix is shifted before applying next convolution operation.
C. Pooling: This operation is used to reduce the spatial size of matrix, thereby reducing cost and complexity. Here values across the sub-matrix of interest are pooled to a single value. Popular pooling techniques include max pooling (maximum value is retained) and average pooling (average value is retained).

Apart from these padding may be used to retain image size and give equal impetus to edge features. A CNN model comprises multiple layers with each layer on right hand side identifying more complex features of an image. For example, the first layer may be identifying primitive shapes like lines, edges etc. While the next layers identify more complex shapes inside the image. Finally at output whole image is identified or classified. One of the first architecture based on CNN model was proposed by as LeNet [4] [5]. LeNet-1 was proposed to identify handwritten US zip codes comprising of digits. LeNet-5 architecture is generalized form of neurocognition. Authors have justified that CNN outperforms all other image recognition models thus far. Different banks utilized LeNet-5 for recognizing handwritten numbers on cheques.

When it comes to learn and make predictions about sequential data including language translation, time series forecasting etc, CNN have certain limitations. CNN are limited to model relation between different data items in sequential dataset. For example, if we want to predict about number of cases of an infectious disease like Covid-19 then we need to investigate recent historical data. The relationship between different data items of dataset cannot be modelled directly into CNN. Apart from this there some other limitations of CNN that make them unsuitable for sequential data. Different sentences input to a language translation model may be of different length but input and output size of CNN architecture is static. These inherent limitations of CNN led to investigation of another deep learning algorithm namely Recurrent Neural Networks (RNN) for processing of sequential data.

III. RECURRENT NEURAL NETWORKS (RNN)

RNN [6] were proposed to overcome the limitations of CNN, inferring sequential data including Natural Language Processing (NLP), Stock Analysis, Biological Sequences, Pattern Matching, Sentiment Analysis etc. As the name suggests they employ neuros with recurrence. To elaborate let a sentence be “My name is abc.” There are four text/words in this sentence which has a definite order. Now, while processing text “name”, RNN must consider the previous output i.e. “My”. To differentiate processing of CNN and RNN (fig. 1) let us write their basic equations:

CNN: $\hat{y} = f(x)$ i.e. predicted output is a function (nonlinear) of input vector, x.

RNN: $\hat{y_t} = f(x_t)$ -(1)

$\hat{h_t} = f(x_{t,}h_{t-1,})$ -(2)

Output of RNN dependents upon the present input and previous output, hence the use of timestamp. The first equation predicts output at timestamp, t which is a function of input at timestamp, t. While the second equation represents the recurrence, hidden layer output at timestamp, t is a function of input at timestamp, t and hidden layer output at timestamp, t-1 i.e. it depends upon the previous prediction. The arrow in fig-1 represents the recurrence. Hence the hidden layers at different timestamps are connected recursively with pervious hidden layers. This enables these models to process the sequential data. Various feature of RNN includes:

A. They can handle variable length sequences.

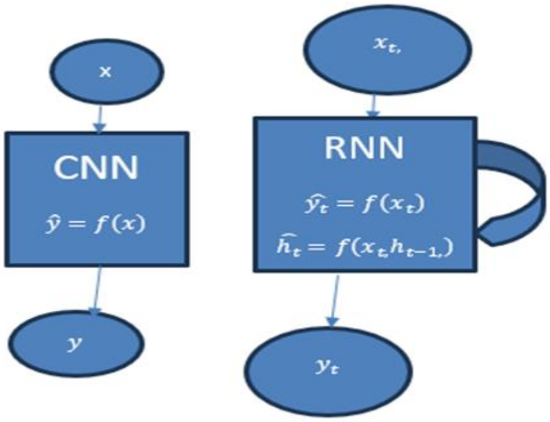

Fig. 1: CNN vs RNN

B. They can trace dependencies among different inputs.
C. They maintain information about orders.
D. They share parameters across the sequence.

Learning in RNN is assisted by BPTT (Back Propagation though Time). Though RNN have inherent inclination towards processing of sequential data, yet they have some limitations:

A. In case initial weights are big then during backpropagation they cause exploding gradient problem. One probable solution to this problem can be clipping gradient.
B. If initial weights are too small, then then they lead to vanishing gradient problem. Which is more concerning than the exploding gradient problem because it makes RNN unsuitable for larger sequences. Though there are some way-around to dealt with vanishing gradient problem like:
   a. Use a more suitable activation function like ReLU in place of tanh or sigmoid.
   b. Rather than initializing weights/parameters to zero, they are assign some initial value.
   c. Modified RNN: Some architectures have been proposed to make RNN suitable for long sequence processing. These make use of gates (GRU: Gated Recurrent Units) to prioritize which information to be retained for long term (future use) and what is to be forgotten. One such very popular architecture is LSTM (Long Short-Term Memory) [7]. It utilizes four gates viz. forget (to forget unimportant information), Store (to store relevant), update (selectively update cell state) and output (return selected values for given timestamp).

There are also some other limitations of RNN that make them unsuitable for large sequences including encoding bottleneck, they do not support parallelism (due to recurrence), no long-term memory, not highly scalable etc.

RNN though are used in some machine learning problems like music generation, sentiment analysis where there is no long-term dependency among input.

IV. TRANSFORMER MODEL

Transformers ever since they were proposed by [8] in 2017 has revolutionized the field of NLP. Especially Large Language Model which were difficult to be realized with its predecessor model RNN have become a reality and huge success. So that so world is now talking about responsible AI and we are ready to enter an arena where machines are ready to replace/substitute humans in their jobs if not humans itself. Government and researchers across the globe are talking about controlling the direction in with AI is heading. And one most important research behind this scenario is the introduction of transformer models. They not only overcome the limitations of RNN in large sequence processing but also have introduced parallelism. Coupled with GPUs that are available now a days they have made it possible to realize large language models like GPT (Generative Pretraining Transformer) by OpenAI and BERT (Bidirectional Encoder Representations from Transformers) by Google. Google has gone one step ahead and introduced BARD based on Transformer-XL architecture. Transformer-XL is modified transformer

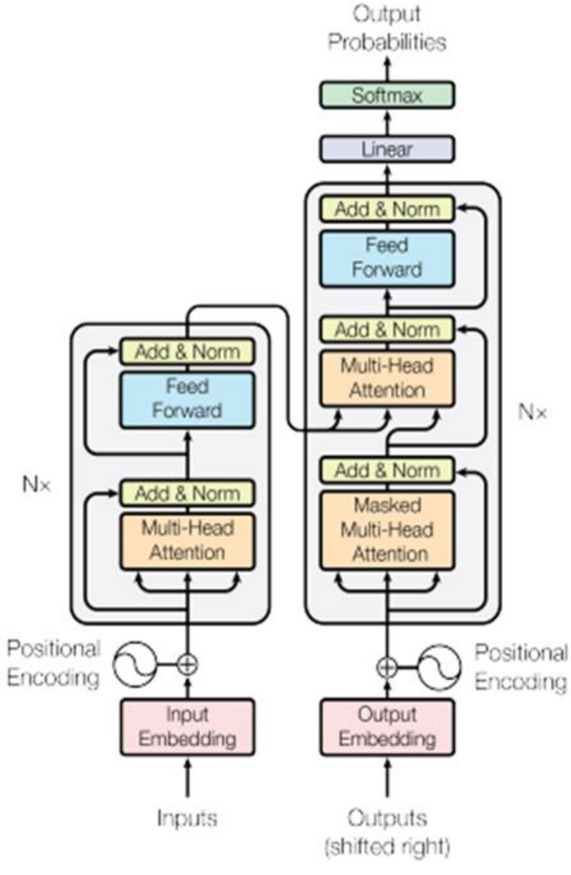

Fig. 2: Transformer Model (Source: https://dl.acm.org/doi/10.5555/3295222.3295349)

architecture to accommodate larger sequences. Original transformer architecture (Fig- 2) comprises stack of two separate units termed as encoder and decoder. Each encoder unit has two sublayers viz. multi-head self-attention and a feedforward network. The decode unit comprises three sublayers viz. masked multi-head self-attention, multi-head self-attention and a feed-froward network. Central to the transformer model is the concept of attention and self-attention. Attention was earlier used two years back in 2015 (before proposal of transformer model in 2017) for image processing. Most image have a central feature that is most attention by the viewer, the same idea was used in concept of self-attention where rather than processing whole image, central or most attended sub-portion was identified. Difference between attention and self-attention is that during self-attention a part of original dataset is focused upon [9].

Attention is represented mathematically as

$$Attention(Q,K,V) = softmax(\frac{Q * K^T}{\sqrt{d_k}}) * V$$

Q (Query), K (Key) and V (Values) are three input vectors. Overall attention tries to quantify similarity between parts (words) of data (large sequence). dk represent dimension of key vector. SoftMax function as usual assign relative probability to each term. Another addition to the transformers is multi-head self-attention. Here attention is applied multiple times for different representations at different positions. Mathematically,

$$Multihead(Q,K,V) = Concat(head_1, head_2\, , \dots \dots \dots, head_h)$$

Here, $head_i$=Attention(QW_i^Q, KW_i^K ,VW_i^v)

W_O is weight matrix as proposed in [8].

So multi-head self-attention essentially help overcome both limitations of RNN architectures viz. recurrence and sequential processing and introduces parallelism into the model. Hence the model is much faster as compared to RNN and can remember long term relations. Another important sublayer of decoding is masked multi-head self-attention, during this all words that proceed the word being processed in the sequence are masked/ignored while only words preceding are accounted for. Positional encoding is used to retain the orders of words in the sequence. Finally, it returns output in term of probabilities.

While predicting next word in a sequence it would propose a number of possible words with relative probabilities. Say for an example the sequence is:

“Man moved out of ________”.

Transformer output may be a set of values like (house: .51, car .39, bus .10).

V. CONCLUSION

During this work three major models of deep learning are reviewed. It was observed that CNN are most suitable and hence most used for image processing in machine learning. Their success can more or less be attributed to the advent GPUs. When it comes to processing of sequential data, they have very limited application. So RNN were used to process the sequential data like NLP, timeseries forecasting, sentiment analysis. RNN too have limited applications while processing large sequence of data due to recurrence, sequential processing and vanishing gradient problems. To overcome these and introduce parallelism in large sequence processing, transformer model introduced during 2017 have made lasting impact. Central to them is the concept of attention. Most attended feature is identified and similarity between words is used to retain information across the large sequence. They are employed in all Large Language Models (LLM) which in turn are revolutionizing the relation between human and machines.

ACKNOWLEDGMENT

Authors feels acknowledged to the members of computer science research fraternity for their contribution in revolutionizing the field of Artificial Intelligence.

REFERENCES

[1] https://openai.com/research/gpts-are-gpts

[2] https://towardsdatascience.com/attention-and-transformer-models-fe667f958378

[3] K O’Shea and R. Nash-*An Introduction to Convolutional Neural Networks,* arXiv preprint arXiv:1511.08458, 2015-arxiv.org

[4] Y LeCun *et al.*, "Backpropagation Applied to Handwritten Zip Code Recognition,"

in *Neural Computation*, vol. 1, no. 4, pp. 541-551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[5] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[6] D Samuel- *A Thorough Review on the Current Advance of Neural Network Structures*, Annual Reviews in Control. 14 (2019) 200-230.

[7] S Hochreiter and J Schmidhuber – *Long Short Term Memory*, Neural Computations, 9(8), 1997, pp 1735-1780.

[8] A Vaswani et. al - *“Attention is all you need”*, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 2017,pp 6000–6010

[9] https://towardsdatascience.com/demystifying-efficient-self-attention-b3de61b9b0fb