Conversational Chatbot with Attention Model

Purushothaman Srikanth, E. Ushitaasree, G. Paavai Anand

Abstract: A Chatbot is an Artificial Intelligence (AI) software that can give a simulation of a conversation between two humans. This Chatbot is based on State of the Art Transformer model architecture which works on Attention mechanism. The transformer model is a very efficient Sequence to Sequence model. Machine translation is at its core, simply a task in which you map the sentence to another sentence. Sentences consist of words that are equivalent to mapping to a different sequence. Beam search and Byte-pair encoding are the algorithms used in our model for heuristic searching in decoder units. A combination of many Unsupervised prediction tasks were carried out by fine-tuning using a multi-task objective every time the user starts the conversation. It takes a new persona for every new session opened and communicates with the persona which is chosen at random. Forwarding the perplexity by the ability to understand and generate natural language this model gives a whooping Hits@1 score efficiency as high as 80.9 percentage.

Keywords: Attention mechanism, transformer, conversation Chatbot, Beam search, Byte-pair encoding

I. INTRODUCTION

In recent times the communication between the machine and human is improving drastically with the growth of AI, which has led to the birth of the Chatbot. Nowadays, the need for an AI assistant is seeing surprising growth. With the usage of Chatbots in almost all websites, specific contents are obtained. Chatbots have improved marketing and customer understanding of the product [0]. Using Deep Learning and Artificial Intelligence we are going to novitate the computers which understand only binary numbers. To react to a person based on what they ask or demand, here Natural Language Processing (NLP) supports and guides this computer machine to understand human language. Which in-turn gives an output which sustenance the user to arrive at a position where they think that answer was on par with human experti.

The technology used here is new and the powerful Attention mechanism architecture which is very effective when it is used for language to language conversion. One type of network built with attention is called a Transformer[10]. The only difference between Recurrent Neural Network and Attention Model is input processing after each word in a sentence is interpreted from the encoder and intermediate levels are generated after every word in the sentence is understood and the collection of these levels is directed to the decoder[6]. The benefit of using the Transformer model is that the sequence of the information is not changed hence the context will be correct.

Beam Search is a heuristic search algorithm that extends the greedy search and returns as many output sequences as possible.

II. LITERATURE SURVEY

Chatbot Using a Knowledge in Database [2]. In that, each of the patterns is paired with the knowledge of Chatbot that was already processed from various sources. The representation and implementation of knowledge in the database table is carried out using SQL (structured query language). The data is modelled on the conversation and the result of it with the Chatbot is examined. Natural Language Processing (NLP) is used which gives the capability of the interaction of machine with the users. There are three NLP analyses: knowledge-based structures, semantic interpretation, and parsing.

Sebastien Jean et al in 2017 [3] presented a technique dependent on significance inspecting which constrains the restrictions of the NMT model by empowering us to utilize an enormous objective dataset without expanding preparing multifaceted nature.

Kyunghyun Cho et al [4] proposed the methodology of a novel Neural Network model called RNN Encoder-Decoder which comprises of two repetitive Neural Networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation and the other decodes it. When a source sequence is given, the encoder and decoder are trained to minimize the conditional likelihood of a target sequence.

The conversational service can provide better interaction and counselling to everyone. It is important to find and resolve issues related to mental disorders such as depression and sadness. A one-to-one conversation can resolve better mental disorders effectively. The work on the cognitive bias modification (CBM) approach has attracted attention.

Chatbot using Gated End to End Memory networks [8] was a research paper written by Jincy Susan Thomas, Seena Thomas in 2018. This intelligent bot system used deep learning method known as Gated end to end memory networks. The bots used Retrieval based model. Incorporated an iterative memory access dynamic regulation of memory interaction. This is used for hospital reservation alone.

Companion Chatbot was a research paper written by Harsh Agrawal et al [7]. Hierarchical recurrent Neural Network (HRNN) is used for better result of specific emotion and shorter training time, since it deals with emotions sentimental analysis is used a
sequence to sequence Neural Network HRNN, it is a reward-based method which assigns some weight to content. It collaborates with the sequence of sentence to give an effective response. It shows all types of emotions like anger, sadness, fear, etc.

Incrementing RASA’s open-source Natural Language Understanding pipeline was a research paper written by Andrew Rafla and Casey Kennington [6]. As open sourced service for NLP is increasing, RASA was updated with incremental intent recognition model as a component to RASA evaluations on the snips dataset show that the RASA to function effectively. Additional features and components are added to RASA (NLU). Only RASA can be used to build the Chatbot to implement these features.

Review on Mood Detection using image processing and Chatbot using Artificial Intelligence[5] was a research paper written by D.S Thosar, et al. The Chatbot can identify the user’s mood based on text or speech using text processing and will provide web links to change the current mood. It can change the mood of the person and give suggestions on what one has to do. It always doesn’t provide customized results and suggestions, so we might not always like the reply.

Attention Is All You Need[9] is the paper written by Ashish Vaswani et al. Machine translation is used in attention architecture with the transformer model which helped us in understanding that encoder-decoder is not the only way to implement the NLP Neural Network speech translation. This paper shows that focus is a powerful and effective way of replacing repeated networks as a modeling method for dependence.

III. METHODOLOGY DESCRIPTION

Shortcomings of the simple encoder and decoder model:
The very common word we see in all NLP text generation models is a sequence to sequence the encoder-decoder model[7] using recurrent Neural Networks that use multiple iterations for processing an intermediate state. A context vector is a single unit of information that is collected from the encoding and which has to be interpreted. The time taken to arrive at the answer is a slow process. Before the Transformer, for both the encoder and the decoder, RNNs were the most commonly used and efficient architecture.

A. ATTENTION MODEL

Attention can’t use output positions to overcome this. The Transformer uses direct encoding of the position applied to the embedding of the data. Attention weight can be measured in a number of ways, but a basic feed-forward neural network was used by the original attention method.

\[ c_t = \sum_i a_{ti} h_i \]  

(1)
The basic mechanism of attention is simply a dot product between the query and the key. However, the size of the dot product tends to grow with the dimensionality of the query and key vectors, so that the Transformer rescales the dot product to prevent it from exploding into massive values.

\[ e_{ti} = f(s_{t-1}, h_i) \]

\[ a_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{T} \exp(e_{tj})} \]  

(2)

B. TRANSFORMER MODEL

The ability to measure the representation of that sequence at various input sequence positions. Instead of using one sweep of attention, the uses multiple “faces” multiple distributions of attention and different output s of one input. It proposes to encode each location and apply the process of focus, to relate two distant terms of both the w.r.t inputs and outputs themselves, which can then be parallelized, thereby improving the learning.
The wonderful thing about the prototypes of the dialog is that you can speak to them. We need to add one item to communicate with our model: a decoder that will construct complete sequences from our model’s next token predictions. Over the past few months, there have been very interesting developments in decoders and I wanted to quickly present them here to get you up-to-date.

C. DATASET
Following Radford et al.’s research, the model is pre-trained on the Books Corpus dataset containing more than 7,000 unpublished books about 800 M words from various genres like Adventure, Fantasy, Romance[11]. Reddit messages, Book corpus, Persona knowledge base is the dataset used. This is quite a lot of data hence the accuracy and efficiency is more. The web scraping has been using flask and beauty soup. Pre-processing was done for the text which was repeated more than once using pandas. Fine-tuning was done after the pre-training phase. Numpy, Torch, Pytorch, Pandas are the few libraries used to built this model.

IV. RESULTS AND DISCUSSION
The model outflanks the current frameworks by a huge edge on the dataset acquiring 51% supreme improvement in perplexity (PPL), 35% supreme improvement in Hits@1 and 13% improvement in F1.

All the more critically, while the model’s hyper-parameters were tuned on the approval set, the exhibition upgrades mean 46% supreme improvement in Hits@1 and 20% improvement in F1.

The perplexity is recognizably low for an open-area language demonstrating task which might be to some extent because of a couple tedious parts of the dataset like the starting articulations toward the start of every discourse (“Hello, how are you?”) and the duplicate instruments from the character sentences.
Conversational Chatbot with Attention Model

V. CONCLUSION

We have fine-tuned the model and extracted more information from the dataset with the persona it has chosen, we made it more and more efficient. We see Attention architecture with the Transformer model as the future of NLP which would overcome the success of recurrent Neural Networks.

We have shown that translation between languages is successful and also machine translation to communicate in one language effectively, such upgrades can be extended to generative tasks like dialog generation that combine many linguistics aspects such as long-range dependencies, common-sense expertise and co-reference resolution modeling, etc. Important future work is nonetheless needed to understand the most ultimate weights and models. We have a good accuracy number about 80.9% currently.

REFERENCES

5. Review on Mood Detection using Image Processing and Chatbot using Artificial Intelligence D.S Thosar, Varsha Gothe, Priyanka Bhokrade, Varsha Sanap
6. Incrementing RASA’s open-source Natural Language Understanding pipeline was a research paper written by Andrew Rafia, Casey Kemington
7. Companion Chatbot was a research paper written by Harsh Agrawal, Arup Patel, Aditya Bholase, G.Paavai Anand
8. Chatbot using gated end to end memory networks was a research paper written by Jincy Susan Thomas, Seena Thomas on Mar 2018
9. Attention is all we need to be written by Ashish Vaswani and co. was published on 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

AUTHORS PROFILE

Purushothaman Srikanth, is currently studying B.Tech Computer Science and Engineering in SRM Institute of Science and Technology, Ramapuram part- Vadapalani Campus, Chennai. He is a Deep Learning aspirant who is doing many Data Science projects for a while. He is a passionate Python developer interested in making impossible things possible with few lines of errors free code.

Email ID: jspurushothaman@gmail.com

E. Ushitaasree, is currently studying B.Tech Computer Science and Engineering in SRM Institute of Science and Technology, Ramapuram part- Vadapalani Campus, Chennai. She is a Quantum Computing Engineer and Deep Learning Engineer who makes wonderful things using her skills in development. Her enthusiasm to learn new things unleashes the phoenix within.

Email ID: ushitasree123@gmail.com

Dr. G. Paavai Anand, is an Assistant Professor at SRM Institute of Science and Technology, Ramapuram part- Vadapalani Campus, Chennai. She had published many research papers in Machine Learning, Text Mining and Genetic Algorithms.

Email ID: pavaigops@yahoo.co.in