Artificial common-sense based inference using Macaw encoder

Mid-semester project report submitted in partial fulfilment of the requirements for the degree of B.Tech. and M.Tech

by

Sreepathy Jayanand.
(Roll No: CED17I038)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
DESIGN AND MANUFACTURING, KANCHEEPURAM

March 2022

Certificate

I, **Sreepathy Jayanand**, with Roll No: **CED17I038** hereby declare that the material presented in the Project Report titled **Artificial common-sense based inference using Macaw encoder** represents original work carried out by me in the **Department of Computer Science and Engineering** at the **Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram** during the year **2021**. With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: March 6 2022 Student's Signature: Sreepathy Jayanand

In my capacity as supervisor of the above-mentioned work, I certify that the work presented in this report is carried out under my supervision, and is worthy of consideration for the requirements of mid-semester project during the period Jan 2022 to Feb 2022.

Advisor's Name: Dr. B Sivaselvan Advisor's Signature: Dr. B Sivaselvan

Abstract

This report documents the literature survey on the different approaches in natural language processing for predicting subsequent sentences given context and relevant options. This report also proposes a novel approach for the same - using an encoder of a generative model specifically MACAW and some additional layers on top, for common-sense inference tasks.

Acknowledgements

I would like to take this opportunity to express my gratitude to Dr B Sivaselvan, Assistant Professor IIITDM K, for his constant support and guidance.

I would also like to express my sincere thanks to Mrs. Mercy Faustina, for her constant feedback and advice to keep me motivated and for providing a sense of direction.

Contents

List of Figures

List of Tables

Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
TL	**T**ransfer **L**earning
LM	**L**anguage **M**odeling
NN	**N**eural **N**etwork

Symbols

Chapter 1

Introduction

1.1 Background

Human language is astoundingly complex and diverse. We express ourselves in infinite ways, both verbally and in writing. Not only are there hundreds of languages and dialects, but within each language is a unique set of grammar and syntax rules, terms and slang. When we write, we often misspell or abbreviate words, or omit punctuation. When we speak, we have regional accents, and we mumble, stutter and borrow terms from other languages.

While supervised and unsupervised learning, and specifically deep learning, are now widely used for modeling human language, there's also a need for syntactic and semantic understanding and domain expertise that are not necessarily present in these machine learning approaches. NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.[8]

1.1.1 Natural Language Processing

Natural language processing or NLP is a field of computer science and AI dealing with the interactions between computers and human language. The machine must be mad capable

of "understanding" the context and even be able to predict next parts of language, also known as language modeling.

1.1.2 Transfer Learning

Transfer Learning or TL is a paradigm in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. This is a really important paradigm specifically in NLP since NLP models tend to be really large and also require large volumes of training data trained over multiple epochs. Training them from scratch would require an enormous amount of resources and time. Transfer learning can help greatly in NLP because in language processing it makes sense to use knowledge learnt from performing tasks on a particular objective and using it on a different objective as a lot of information can be transferred. Transfer learning is categorised into mainly 3 types.

- Positive Transfer - When learning in one situation facilitates learning in another situation, it is known as a positive transfer. For example, skills in playing the violin facilitate learning to play the piano. Knowledge of mathematics facilitates to learn physics in a better way. Driving a scooter facilitates driving a motorbike.
- Negative Transfer - When learning of one task makes the learning of another task harder- it is known as a negative transfer. For example, speaking Telugu hindering the learning of Malayalam.
- Neutral Transfer - When learning of one activity neither facilitates or hinders the learning of another task, it is a case of neutral transfer. It is also called as zero transfer.

1.1.3 Language Modeling

Language modeling or LM is the use of various methods to determine the probabilities of finding out the subsequent part of language given some contextual information. Some of the major domains that use language modeling is translation, question answering etc.

Usually they are treated as regressive tasks wherein we use the information of the context and the language generated so far to obtain the next parts of language rather than just using the context given to us.

Language modeling can help in simplifying a wide variety of day to day tasks, for example speech recognition. Smart speakers, such as Alexa, use automatic speech recognition (ASR) mechanisms for translating the speech into text. It translates the spoken words into text and between this translation, the ASR mechanism analyzes the intent/sentiments of the user by differentiating between the words. For example, analyzing homophone phrases such as “Let her” or “Letter”, “But her” “Butter”. Machine translation is also a crucial part of language modeling.[9]

1.2 Motivation

Natural language processing is a very important field in the life of humans now adays, especially after the newer advent of models such as the transformers have helped increased the scores in various benchmarks close to that of humans. There are a wide variety of tasks within NLP that can be looked at and further improved. One of the most important tasks is in the discriminative fields within NLP. In this field of NLP, most of the tasks involve dealing with language in a discrete sense. For example given a sentence, find the possible next sentence given ”n” options; Or given a context find the summary of the context; Or given a chat find the sentiment from a list of common sentiments which the chat belongs to etc.

Research in natural language understanding and textual inference has advanced considerably in recent years, resulting in powerful models that are able to read and understand texts, even outperforming humans in some cases. However, it remains challenging to answer questions that go beyond the texts themselves, requiring the use of additional commonsense knowledge.

1.3 Objective

Common sense inference is an aspect when improved upon by machines can lead to significant improvement for machines to bridge the gap between humans. In this report we try to propose a method wherein we can create a model which can understand context from a sentence and be able to predict logical next sentences from the given options.

Chapter 2

Methodology

There are a variety of methods which yield good results in the fields of non-autoregressive tasks. Let us look at some of them

2.1 Transformers

A transformer is a deep learning model that adopts the mechanism of self-attention to convert one sequence to another, or in other words it is a sequence to sequence model. The model consists of an encoder and a decoder. The encoder takes the input sequence and converts into embedding of a particular dimension. This abstract embeddings are taken forwards by the decoder to obtain an output sequence. This particular method can be used of machine translation, summarization etc.[1]

2.1.1 BERT

Bidirectional Encoder Representations from Transformers or BERT[2] is a transformer based model developed by Google for NLP application that is still widely used. It had created state of the art results when it was launched in variety of NLP tasks such as question answering, named entity recognition, natural language inference etc. Usually

BERT models and the ones derived from BERT usually output a class label or a span on text.

There are other generative models such as T5, which has better capabilities for generative tasks.

2.1.2 T5

Text-to-Text Transfer Transformer or T5[3] aims to convert any and all task to generative tasks and can give satisfactory output for a variety of tasks such as machine translation, classification tasks, regression tasks etc. T5 was trained on the common web crawl or c4 dataset to train it on 750GB of clean English text. The T5 model uses the transformers as the building blocks with each encoder block containing a self-attention layers and a feed forward layer.

2.2 Solving common-sense inference tasks

In order to solve various commonsense inference tasks we can go with any of the above mentioned popular methods. Obtaining abstract high dimensional representations of tokens given in the input text is the most difficult part. This part is natural for the usage of BERT as it is training for generative powerful embeddings. We can try exploring the embedding powers of T5, by only using the encoder of T5 rather than using T5 as a generation model as described in ENCT5 [4].

Chapter 3

Literature Review

Let us look at some of the different approches and models at tackling the tasks which we have discussed.

3.1 RNNs

Recurrent Neural Networks or RNNs is an machine learning model ideal for sequential data such as text, financial data, speech, audio, video etc. RNNs virtually have an internal memory that allows the previous inputs to affect the subsequent predictions. This would particularly be a really good strategy for language generation as the next word is heavily dependent on the previous words.

3.2 LSTMs

Long-short term memory or LSTM are a betterment of RNNs but is better than traditional RNNs having a better memory compared to the latter. During the multiple learning processes of LSTMs, it learns to retain relevant information and lose the irrelevant information. This is done using forget gates, which is responsible for calculating the cell state which is not relevant so that it can be discarded.

3.3 Siamese NNs

Siamese neural networks for natural language classification are 2 similar neural networks each capable of generating embeddings. After obtaining the different embeddings that can be taken and various similarity measuring functions can be used to find the "distances" between various candidate answers and the question.

3.4 BERT

BERT[2] for sequence classification is a model which can help in classifying the given input to a set of given labels. This is done by adding some fully connected layers on top of the BERT encoder to try to tune it down to the number of tokens.

3.5 DialoGPT

Dialogue Generative pre-trained transformer or DialoGPT[6] is a generative model trained on 147M conversation-like exchanges extracted from Reddit comment chains. It can attain the performance close to humans in signle-turn dialogue settings. It is built on top of GPT-2. It uses a multi-layer self-attentive mechanism to allow fully-connected cross attention to the full context in a computationally efficient manner.

3.6 T5

T5[3] is a generative model trained on the colossal clean crawled corpus or C4 dataset. It achieves the state-of-the-art results on many generative benchmarks. It can be used in a variety of tasks such as machine translation, summarization, question answering, classification. Even the classification task is treated as a generation task, which the whole point of T5, i.e every task is treated as a generation task. In the comprehensive paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" it is discussed how the encoder-decoder architecture prevails the existing decoder only

architecture. It is trained not only in one task but in multiple tasks thus making it a one model fit all task approach. The encoder of T5 can be used to obtain the embeddings and we can add another linear and dropout layers on top to tune it down to obtain the labels possible for each text instance.

3.7 EncT5

Encoder-decoder transformers are some of the most promising architectures for generative tasks. But they can also be used for other tasks effectively. It is also more favourable over architectures such as BERT for pre-training on language model task when it comes to large scale models which could take months to train given it's generality. This paper "EncT5"[4] proposes a way to efficiently fine-tune pre-trained encoder-decoder models such as T5 for classification and regressive tasks using the encoder layers alone and leaving out the decoder layers. Hence with less than half the parameters of T5 it can perform similarly to T5 on various benchmarks.

3.8 MACAW

MACAW[5] is a versatile, generative question-answering (QA) system that is built on UnifiedQA, itself built on T5 and exhibits strong performance in a wide variety of tasks, and outperforms GPT-3 in the "Challenge 300", a suite of 300 challenge questions, despite being an order of magnitude smaller. The MACAW large has only 11 billion parameters compared to 175 billion parameters of GPT-3.

MACAW allows different permutations or angles, take a question get the answer; take an answer and obtain a suitable question; or take a question and answer, and obtain a series of options.

3.8.1 Why use MACAW?

Macaw is a model capable of general question and has been built on top of UnifiedQA[7], and further fine tuned on the ARC and ARC-DA datasets, which are grade school-level, MCQs to encourage research in advanced question-answering. This makes Macaw a good model to further work on for common-sense inference tasks as well. Here is an example from ARC dataset:


```
{
  "id": "MCAS_2000_4_6",
  "question": {
    "stem": "Which technology was developed most recently?",
    "choices": [
      {
        "text": "cellular telephone",
        "label": "A"
      },
      {
        "text": "television",
        "label": "B"
      },
      {
        "text": "refrigerator",
        "label": "C"
      },
      {
        "text": "airplane",
        "label": "D"
      }
    ]
  },
  "answerKey": "A"
}
```


3.8.2 What is Macaw capable of?

As we've seen, Macaw was trained on the ARC dataset which makes it pretty good at question answering in general which can be further fine tuned for common sense inference. Let us look at some sexample that Macaw generated[5]:

Q: A young boy was rushed to the hospital emergency room, but the ER doctor saw the

boy and refused to operate. "This boy is my son," the doctor said. But the doctor wasn't the boy's father. How could this be?

A: The doctor mistook the boy for his own son

Q: How would you make a house conduct electricity?

A: Paint it with a metal paint

Chapter 4

Problem statement and proposed solution

4.1 Problem Statement

Given a context/sentence and a list of possible sentences which can follow the given context logically, find out the best possible follow on sentence from the list of options.

4.2 Dataset

The Macaw model was fine tuned on ARC dataset, a high school based science question and answer dataset. It has the ability to reach logical conclusions from being trained on this dataset. The dataset chosen for this project is the Situations With Adversarial Generations or SWAG dataset, which contains logical follow on sentences from a given source sentence although it is not specifically a question and answer dataset. The dataset contains descriptions like "she opened the hood of the car," and we are expected to reason out the situation and anticipate what might come next, eg - "then, she examined the engine".

4.3 Proposed solution

The EncT5 paper promises pretty good performance from the encoder of an encoder-decoder model. Macaw is a T5 based model, which itself is an encoder-decoder model, specifically trained for reasoning based question and answering. This makes for a good case that the encoder of the Macaw model can be further used elsewhere. So the proposed model would involve the encoder of the MACAW model(encoder-decoder generative model) and added on top of this would be a series of linear layers, outputting the probabilities for the various follow on sentences.

Chapter 5

Extensions and future scopes

The proposed model can be adjusted to support any number of labels at the output. So we can make it work for various other tasks such as question answering, sentiment analysis etc, given we have an appropriate dataset to fine tune the model on. Since the Macaw encoder is pre-trained on tasks involving general reasoning, it can be expected to work reasonably well in many of the tasks.

Bibliography

[1] "Vaswani et. al. - Attention Is All You Need (June 2017)" [Online]. Available: https://arxiv.org/abs/1706.03762

[2] "Jacob Devlin et. al. - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Oct 2018)tra" [Online]. Available: https://arxiv.org/abs/1810.04805

[3] "Colin Raffel et. al. - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (Oct 2019)" [Online]. Available: https://arxiv.org/abs/1910.10683

[4] "Frederick Liu et. al. - EncT5: Fine-tuning T5 Encoder for Non-autoregressive Tasks (Oct 2021)" [Online]. Available: https://arxiv.org/abs/2110.08426

[5] "Oyvind Tafjord et. al. - General-Purpose Question-Answering with Macaw (Sept 2021)" [Online]. Available: https://arxiv.org/abs/2109.02593

[6] "Yizhe Zhang et. al. - DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation (Nov 2019)" [Online]. Available: https://arxiv.org/abs/1911.00536

[7] "Daniel Khashabi et. al. - UnifiedQA: Crossing Format Boundaries With a Single QA System (May 2020)" [Online]. Available: https://arxiv.org/abs/2005.00700

[8] "Transfer of Learning: Types and Theories" [Online]. Available: https://www.trainerslibrary.org/transfer-of-learning-types-and-theories/

[9] "Language Models" [Online]. Available: https://insights.daffodilsw.com/blog/what-are-language-models-in-nlp

An approach for commonsense inference using Macaw encoder
Mid-Sem Project Presentation

Guide: Dr. B Sivaselvan

Asst Prof. IIITDM

Sreepathy Jayanand

CED17I038

Problem Statement:

- Common-sense inference is a task in natural language processing involving understanding the context and suggesting possible logical extensions from the context.
- E.g. :“A young boy was rushed to the hospital emergency room, but the ER doctor saw the boy and refused to operate. ”This boy is my son,” the doctor said. But the doctor wasn’t the boy’s father. How could this be?”
- Possible answer: The doctor is boy’s mother.
- Macaw’s answer: The doctor mistook the boy for his own son
- The problem that we are trying to solve here is – Given a context and a list of possible extensions from the context, select the most probable option.

Literature Review:

- Here are 2 of some of important papers leading to the proposed solution.
  - Multi angle question answering - MACAW
  - EncT5: Fine-tuning T5 Encoder for non-autoregressive Tasks

MACAW:

- Macaw is a generative question answering model that is built on T5, and only has 11billion parameters, as compared to GPT-3, a leading generative model which contains 175billion parameters but still produces similar benchmarks.
- T5 is an encoder-decoder generative model that has achieved the state-of-the-art benchmarks in a variety of generation tasks such as language translation, summarization, question answering etc.
- Macaw allows different “angles” – For example, Given a question, it can generate the answer; Given an answer it can generate possible options; Given the option it can generate a question etc.

MACAW:

- It is fine-tuned on the ARC dataset, which is a school level science-based reasoning dataset with a question with multiple options.
- Hence, we can expect it to learnt a good context in general common-sense based questions.
- E.g. Q: How would you make a house conduct electricity?
- Macaw's output: Paint it with a metal paint
- Macaw's syntax:
  - "$question$ = What is the color of a cloudy sky? ; $answer$; $mcoptions$"
  - Macaw's output: '$answer$ = gray ; $mcoptions$ = (A) blue (B) white (C) grey (D) white'

EncT5:

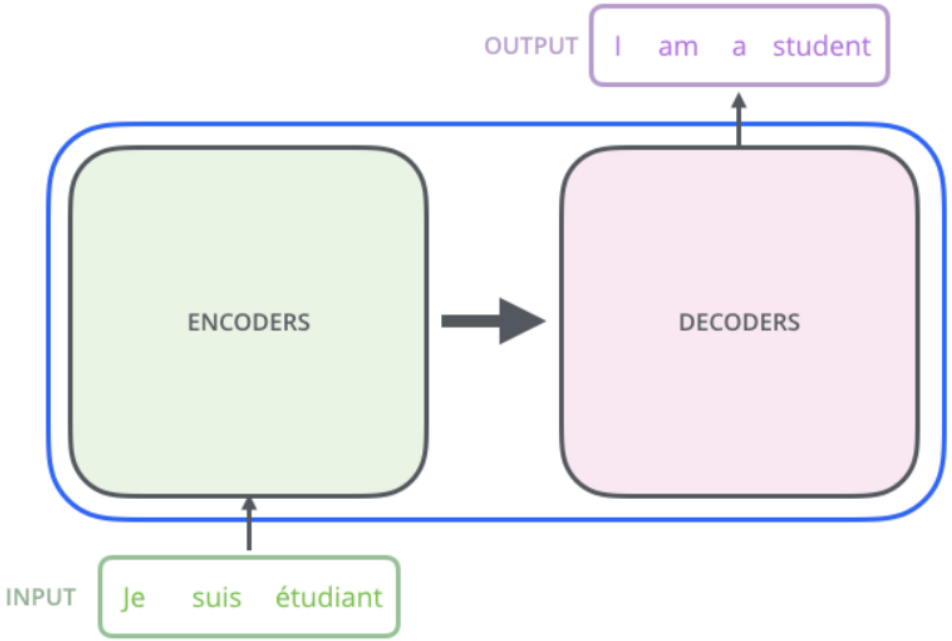
- T5 is an encoder-decoder generative model.
- This paper sheds light on the fact that the encoder of a trained T5 model can be used to obtain embeddings of the input sentence which would have more representational power.
- These encoded input vectors would have good representational power of the input sentence and can be used further.

SWAG dataset:

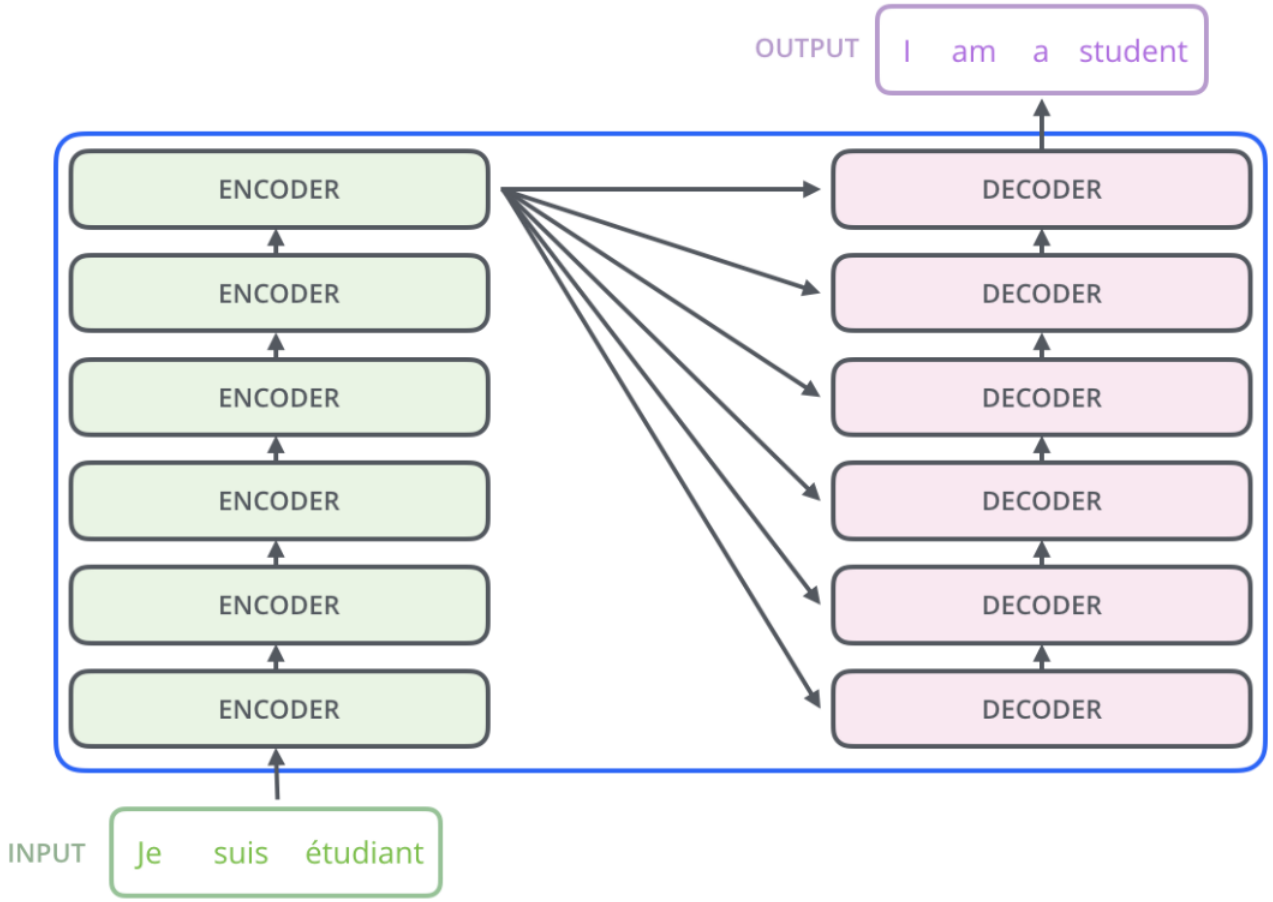
- SWAG / Situations With Adversarial Generations dataset contains instances where in a situation is present and a good amount of “common-sense” is required for predicting the next phase of the situation.
- E.g. “On stage, a woman takes a seat at the piano. She,
  - A. sits on a bench as her sister plays with the doll.
  - B. smiles with someone as the music plays.
  - C. is in the crowd, watching the dancers.
  - D. nervously sets her fingers on the keys – Correct option

Architecture of Macaw:

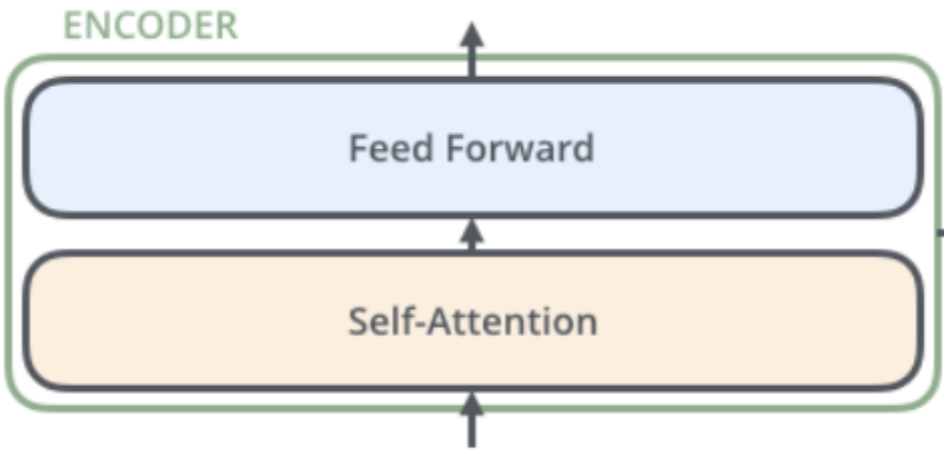
- Macaw is essentially and encoder-decoder architecture, like the transformer.



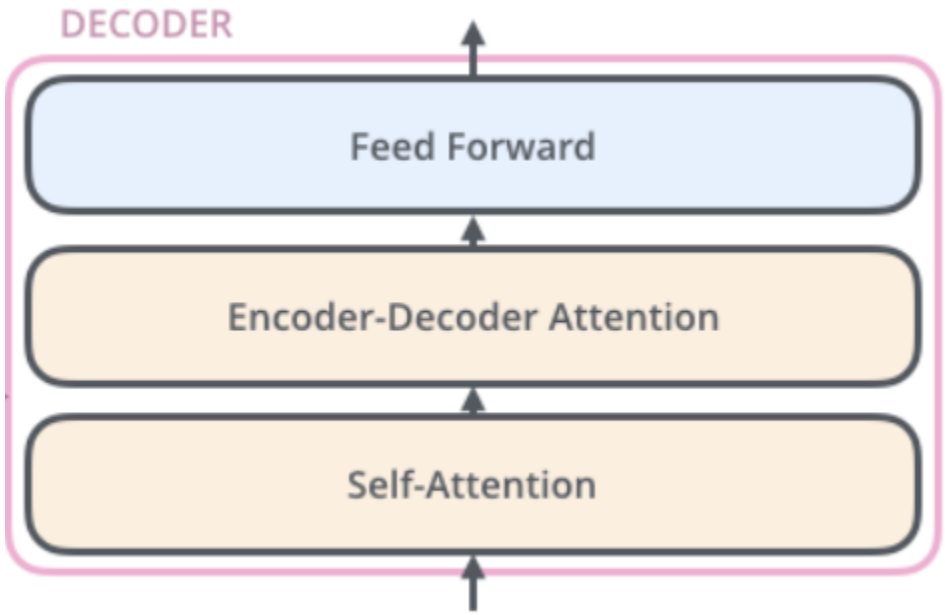
Architecture of Macaw:



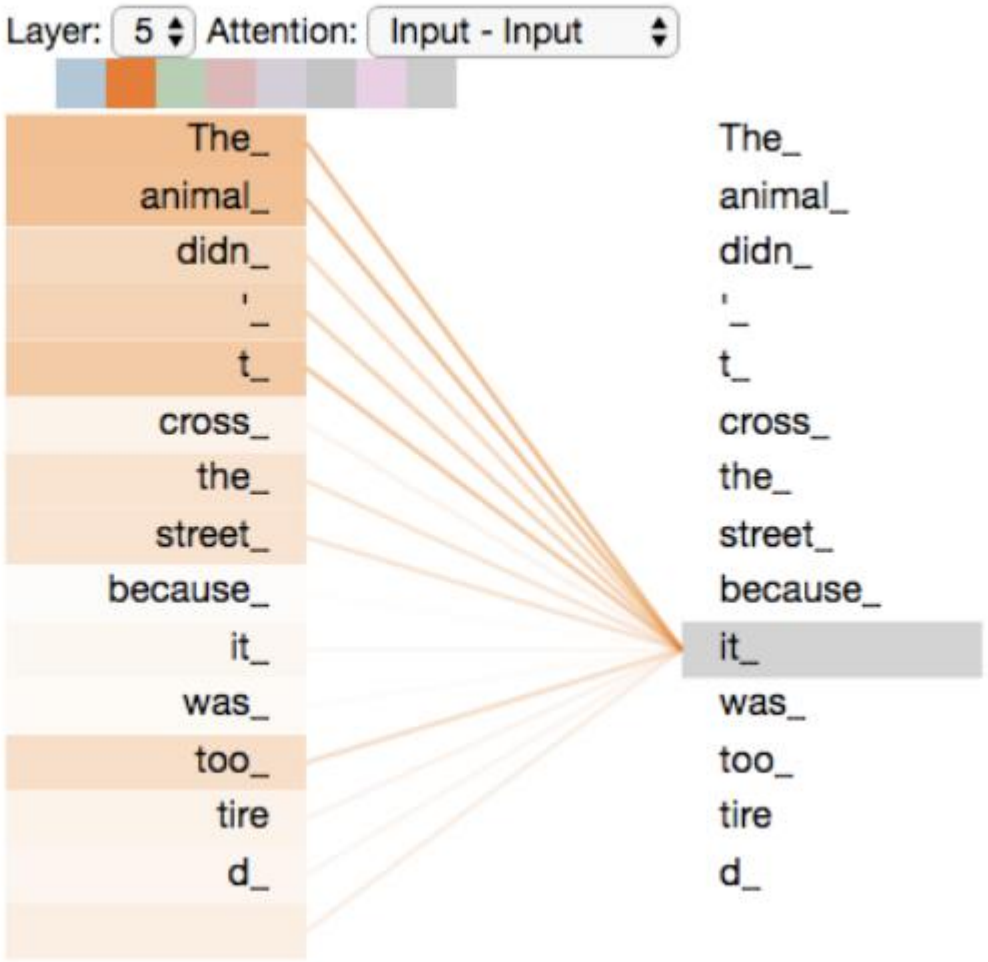
Architecture of Macaw:



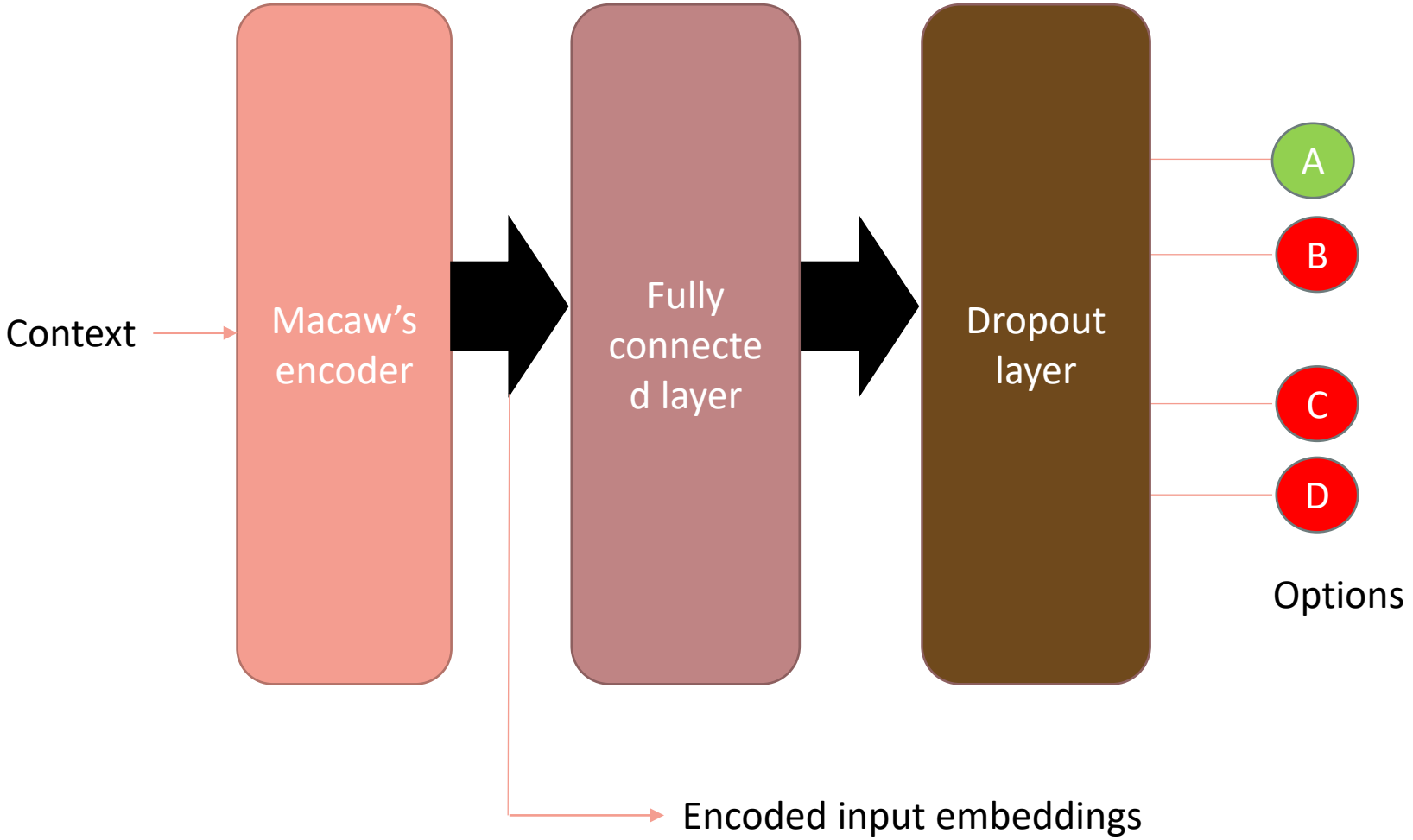
Architecture of Macaw:



Architecture of Macaw:



Proposed Solution:



THANK YOU