Develop New Transformer Architecture For  
Question and Answering

Nirbhay P. Tandon

Attention-based Transformer architectures have become the norm of current Natural Language Processing applications. Google began this trend back in 2017 with their paper *Attention Is All You Need*, by introducing the Transformer architecture that works solely on attention mechanisms. The purpose of our work will be to create a new kind of Transformer architecture. Compare and contrast its performance against other architectures such as BERT, DistilBERT, ALBERT etc. using the SQuAD 2.0 Dataset. Through our research, we aim to produce a new architecture that has a better performance than existing models specifically for Question Answering based applications.

# Introduction

One of the most philosophical questions that anyone can encounter is *“Can an answer exist without a question?"*. While that question exists to be solved by philosophers and thinkers, we shall aim to address the *answer* that can be provided to a question in terms of the context that has been posed. While the field of Natural Language Processing(NLP) is vast and many unexplored areas exist, in this thesis the focus is entirely on developing a novel architecture that provides *highly contextual and relevant answers* to the questions asked of it.

The earliest examples of Question-Answering(QA) systems can be traced back to 1961, in the form of BASEBALL,, a system developed by Bert F. Green and team to answer basic questions posed using punch cards in ordinary English language about stored data on baseball games. The team implemented a dictionary structure to allow for answers to be looked up and printed. For example, the system could answer very direct questions such as: “Where did each team play on July 7?". A highly impressive feat for the early days of neural networks in 1960’s, the work done by Green paved the way for systems like LUNAR,, answered geological questions for the rocks brought back from the Apollo moon missions. LUNAR, which was first presented in 1971, had an astounding 90% accuracy rate for questions that were asked by people who didn’t know how the system worked! Question-Answering based systems have gained a lot of popularity, especially in the form of “chatbots”. These systems depend highly on contextual understanding of the input, the training corpus and the question asked. They use this knowledge to output an answer that can help the user with whatever their query is. Recurrent neural networks and architectures based on them, have been able to provide great advancements in the field of Question-answering and chatbots in general. However, there is a behaviour of over-fitting and lack of contextual understanding of the question. This, coupled with long training times and extremely complex mathematical model designs, have often kept the field of Natural Language Processing slightly obscured from the masses.

The problem of Question-Answering is of a sequential nature. A sequential problem is one which

In a more mathematical context, we can say that the problem we face is of *answer selection* and not *answer provision*. Through numerous studies that have been done, some of which we shall look into detail in chapter [[c2]](#c2), we see that the problem is not how to provide an answer to a question, it is if the answer provided is the correct one or not in the *context* in which the question has been posed.

Through the work done in this research, the author aims at creating a sustainable, fast, easy to understand Transformer architecture for Question Answering. An advancement on the work done by the team at Google . To be able to do so, let us first understand what a *Transformer* is. A Transformer is a form of transduction model that relies solely on self-attention to figure out how to represent its inputs and outputs. It does so without the use of any sequence aligned recurrent neural networks(RNNs) or convolutions.

## Background Of The Study

In this section, we shall highlight what has led us this far and some of the interesting challenges that it poses. In [[back]](#back), we briefly look at the history of Natural Language Processing and how some of the challenges were addressed. In [[rr]](#rr), we look at the latest research that has gone into creating the Transformer architecture, identify some of the common patterns and use that information to strategise our model in later sections.

The area of Natural language Processing has taken significant leaps in the last two decades. Work done towards improving the ability of machine learning models to first recognize words, then sentences, followed by contextual understanding has led to several interesting and novel approaches in the field. From early on neural networks to creating Long Short-Term Memory architectures by Sepp Hochreiter and Jurgen Schmidhuber in the mid-’90s that resolved the vanishing gradient problem of classical neural networks, we have come a long way.

The latest advancements in this field come from Google’s research lab in the form of *Transformers*. We look into this in a bit more detail later. However, no model can be successful without a good dataset to train on. This is where the SQuAD 2.0 dataset comes in. This dataset is what forces the machine learning models to do contextual understanding. One might even say that it forces the models to “think” for themselves before answering a task.

Let us now look at some of the key research that has been done in this regard.

## Aims And Objectives

The main aim of this research is to propose a new transformer architecture that can perform better at Q&A using the SQuAD 2.0 dataset. We shall:

1. Implement the existing models that are available via libraries such as HuggingFace , PyTorch and Tensorflow on the dataset
2. Obtain F1, Exact Match(EM), etc. scores for existing models and treat them as our benchmark scores
3. Identify drawbacks of the current architectures
4. Design our architecture and evaluate its performance
5. Fine-tune the architecture, re-evaluate and report improvements
6. Compare the results of our Transformer model with the benchmark scores.

## Scope Of The Study

This research work is entirely focused in the area of Question-Answering based networks. This focus helps in significantly narrowing down the scope and critically assessing the problem that needs to be solved. The practical applications of the field of QA are endless. From chatbots to QA systems on domain specific knowledge or conversational applications for voice assistants, a light, robust, highly accurate Transformer based architecture can help ease the daily lives of millions of users.

## Significance Of The Study

The current state of the art systems, some of which have been highlighted in section [1.1](#c1background), suffer from various drawbacks and even the best models have only been able to achieve an EM score of 90.871 with an F1 score of 93.183 on the SQuAD 2.0 leaderboard .

## Structure Of The Study

We have divided this report into 3 main chapters. Each chapter is broken down into sections and subsections to localize the reference and use it further in the global context of this research. In chapter [1](#c1introduction), an introduction to the concept of Question-Answering, a brief look at the problem, the motivation, the scope of this research and its significance have been provided. Chapter [2](#c2litrev), looks in detail at the background studies that have been briefly mentioned in section [1.1](#c1background). In this chapter critical analysis is conducted how some of the work that has been selected is relevant to what this research aims to highlight, it helps us understand the merits and shortcomings of current state of the art approaches. Finally, Chapter [3](#c3researchmeth), details the research methodology used by the author. Chapter [3](#c3researchmeth) helps the reader understand various nuances of the dataset, evaluation metrics, pre-processing techniques, benchmarks from existing systems etc.

# Literature Review

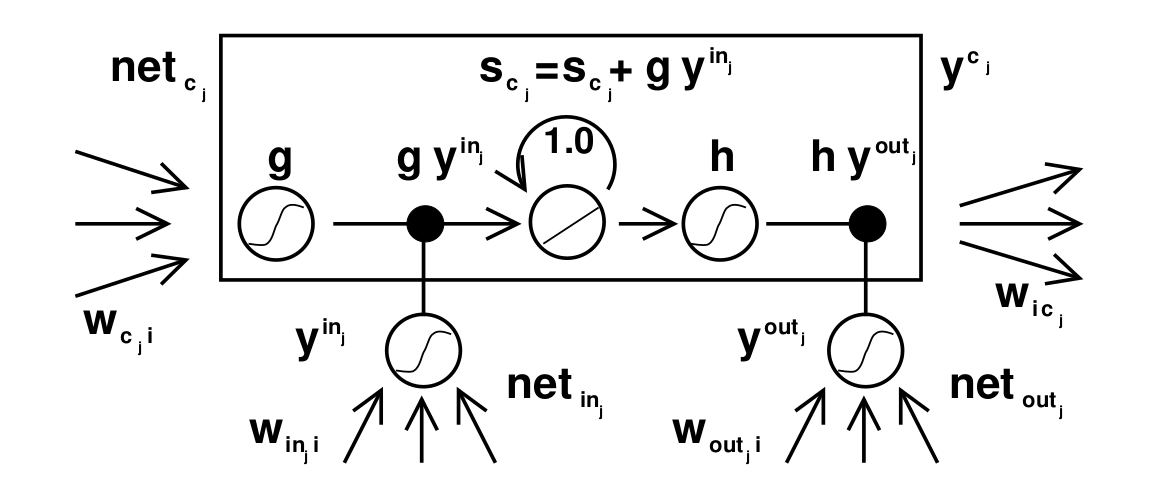
To be able to effectively deliver on the aims and objectives we have defined in Section [1.2](#Xb52009b64fd0a2a49e6d8a939753077792b0554), we must first critically analyse the work that has been done in the field of NLP and, in particular, the field of Question-Answering based networks. In this section, we will critically analyse a few key model architectures and how we reached Transformers, which form the base of our proposed architecture.

## Long Short-Term Memory

In 1997 introduced to the world a gradient descent based model in the form of Long Short-Term Memory. They set out to solve the problem of the vanishing gradient which was often seen in "Back-Propagation Through Time"(BPPTs) based neural networks. Extensive studies done on this, some by Hochreiter himself, showed that the problem of vanishing gradients is a real one. It was also seen that in case of BPTTs, the error back-flow mechanisms would either blow up or also suffer from vanishing gradients leading to either oscillating weights or learning to bridge long time gaps would not work. To remediate this, the authors introduced the concept of a *constant* error flow through the internal states of special units.

Their paper, Long Short-Term Memory, , highlights that the work previously done using various gradient-descent variants, time-delays, time constants, use of higher order units to influence connections when the network receives conflicting error signals, Kalman filters, Second order nets etc. still suffer from some of the same problems that we have earlier. In the case of Ring’s approach , the network can be really fast but also suffers from memory constraints and lack of generalization. This happens because to bridge time lags, the model adds additional units.

It is essential for us to understand how LSTMs function as it forms the basis of Transformers to exist. Hochreiter and Scmidhuber designed a *memory cell* that allows for a constant error flow. This is achieved by introducing *constant error carousel*(CEC), a multiplicative input gate unit and a multiplicative output gate unit. These *multiplicative* units help shield the contents of the memory cell from irrelevant inputs. The CEC, input and output gates together form the *memory cell*, as seen in Fig. [2.1](#lstmCEC).



LSTM Memory cell with a Constant Error Carousel having fixed weight 1.0,

The central feature for LSTM’s memory cell is the CEC. The self-recurrent connection, where weight is 1.0, is what helps create a time-delayed feedback loop by 1.

The CEC uses gate units to circumvent the issue with conflicting weights in the input and output layers. The input layer has control over when to use and when to override the information in the CEC but has no control over the error signals that are in the memory cell. The output gate, however, uses a feedback mechanism to assess when to superimpose different error signals. It also falls on the output gate to *learn* which errors to trap in the CEC by appropriately scaling them. The gates have to learn when to release and trap errors therefore controlling the access to the CEC, helping maintain constant error flow. To mathematically understand the memory cell from fig. [2.1](#lstmCEC), Schmidhuber and Hochreiter proposed the following set of equations [[eq1]](#eq1) - [[eq3]](#eq3). More details on this can be found by referring to the appendix section A.1 in .

where

and

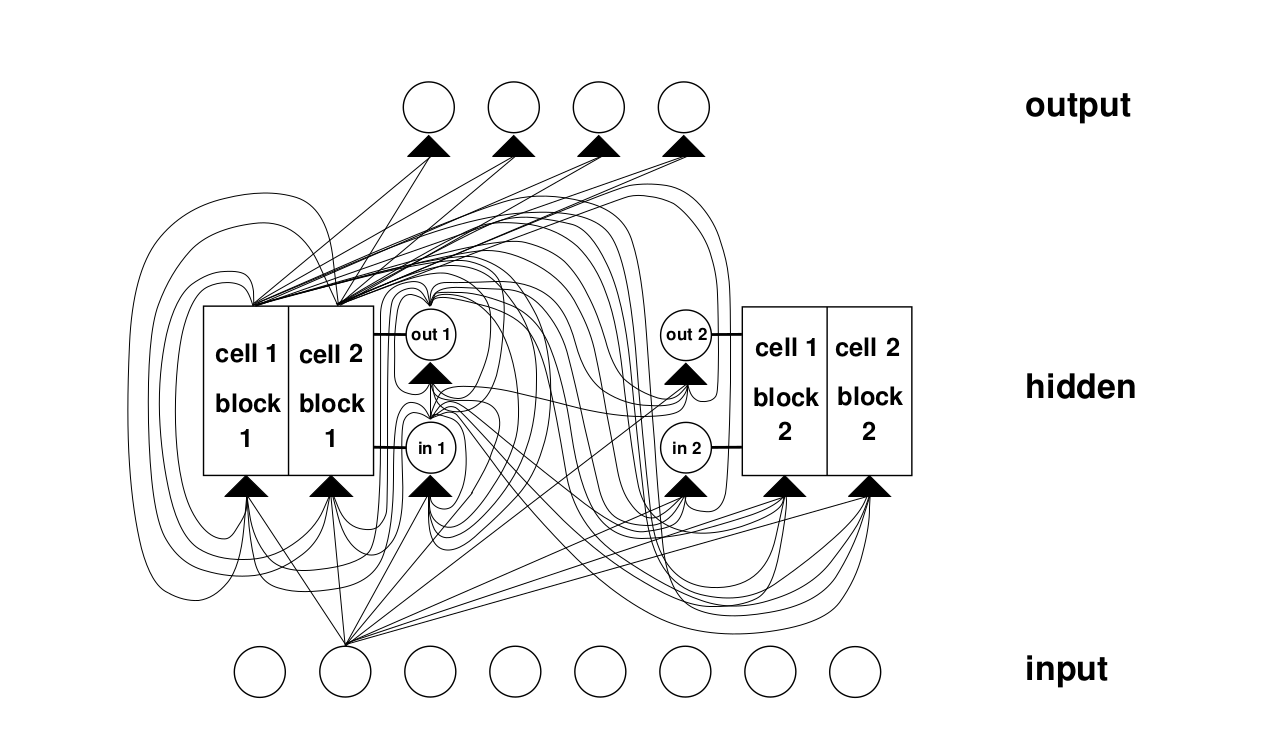
A generalised form of this equation can be represented as:

At time , ’s output is computed as:

where the “internal state” is:

Where may stand for input, output or gate units, memory cells or even conventional hidden units, if they are being used.

In fig. [2.2](#lstmNetwork), we see the internal workings of how an LSTM network could look like.



LSTM network with 8 input cells, 4 output cells and 2 memory cells of block size 2. Here marks the input gate, marks the output gate and marks the first memory cell block 1. The internal architecture of is similar to fig. [2.1](#lstmCEC).

A series of experiments were performed to prove that LSTMs are indeed “state-of-the-art" and provide an improvement over then existing RNN based architectures. Experiments around temporal order, which is defined as a series of events that happen in a particular order, that were previously unsolved by RNNs, were successfully solved by LSTMs. This was possible because of how LSTMs address time lags vs. how RNNs suffer from vanishing gradients. Details of the analysis and experiments conducted are available in section 5 of the paper “Long Short-Term Memory" by , with experimental summaries in tables 10 and 11 of the same. Being able to solve for temporal order based problems allowed LSTMs to be capable of solving various real world problems that we will discuss subsequently. Temporal orders are key in contextual understanding for NLP based tasks and especially in the case of QA based networks. Let us now look at some advantages and limitations of LSTMs, before moving on to understand how they have been applied in the field of QA.

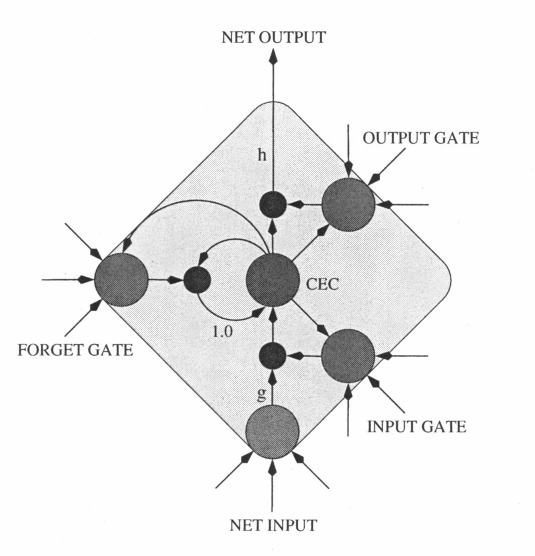
#### Advantages of LSTMs

* LSTMs deal with long time lags by bridging the gaps using constant error flow in the memory cells.
* LSTMs respond well to generalization problems, even if the input sequences are widely separated.
* They work well with a variety of braod range hyper-parameters and don’t usually require tuning.
* Their time and weight update complexity is the same as BPTT, ie. . The advantage for LSTMs however, is that it is local in both space and time.

#### Limitations of LSTMs

* LSTMs use memory cells which require an additional input and output unit. This *hidden unit* is replaced by 3 units in an LSTM architecture, compared to 9 in an RNN. A fully connected LSTM will have connections however.
* LSTMs dont work well when they get the entire input in one go. The architecture is unable to generalize well by random weight guessing in this case.
* They do not have the ability to “natively" count discrete time steps which might be useful in some applications. This isn’t a big issue for the case of QA based problems.

While LSTMs have found various applications across the field of machine and deep learning, QA seems to be one where being able to achieve consistent results of accuracy as well as having an exact match seems to elusive. Research carried out at IBM by Tan, Santos and co. looked at approaching this problem by using something called Bidirectional LSTMs or BiLSTM, . Introduced by Graves and Schmidhuber in their paper *Framewise Phoneme Classification with Bidirectional LSTM Networks*, the biLSTM implements a *forget gate* as seen in fig. [2.3](#lstmBilstm). An advantage of biLSTMs over traditional LSTMs is that you feed the data twice, once from the beginning to the end and then from the end to the beginning. This helps the model contextualize and adjust weights better, therefore creating a more robust model that learns and performs better overall when compared to traditional LSTMs and RNNs.



A memory cell for a biLSTM highlighting a *forget gate*. This forget gate helps in scaling the internal state of a memory cell, for example by resetting the state to 0. We can see that it is different to fig. [2.1](#lstmCEC), with the implementation of a connection to the forget gate, while still having an internal weight of 1.

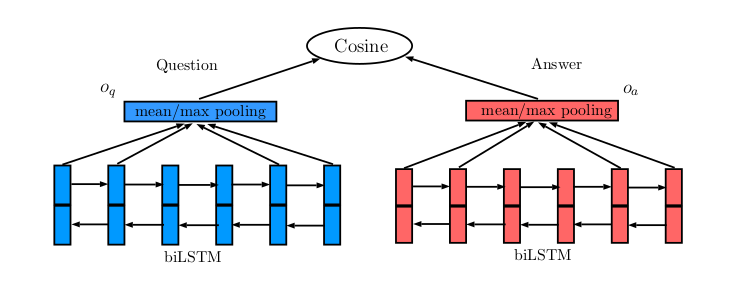
BiLSTMs were introduced as an approach towards phoneme classification for speech recognition applications, however, in their paper “LSTM-BASED DEEP LEARNING MODELS FOR NON -FACTOID ANSWER SELECTION", have shown that extending the biLSTM model 2 different approaches for QA has its advantages. One approach, using a more composite method of using CNNs for representing questions as well as answers. The other, by implementing an *attention mechanism* that helps in generating an answer representation according to the question context. Their framework is based on developing a biLSTM model for both questions and answers, applying a connective pooling layer and using a similarity metric to measure the degree of matching answers. They utilized the InsuranceQA dataset that was created in 2015 and tested primarily on CNN based networks, details of the dataset can be found in table [2.1](#lstmInsuranceQATable).

InsuranceQA Corpus statistics: first two columns are the question and answer quantity; notice there could be multiple answers for some questions so that the answer quantity is larger than the question quantity; third column is the question total word count. The total number of answers is 24 981 and the whole answer text contains 2 386 749 words.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Questions | Answers | Question Word Count |
| Train | 12887 | 18540 | 92095 |
| Dev | 1000 | 1454 | 7158 |
| Test1 | 1800 | 2616 | 12893 |
| Test2 | 1800 | 2593 | 12905 |

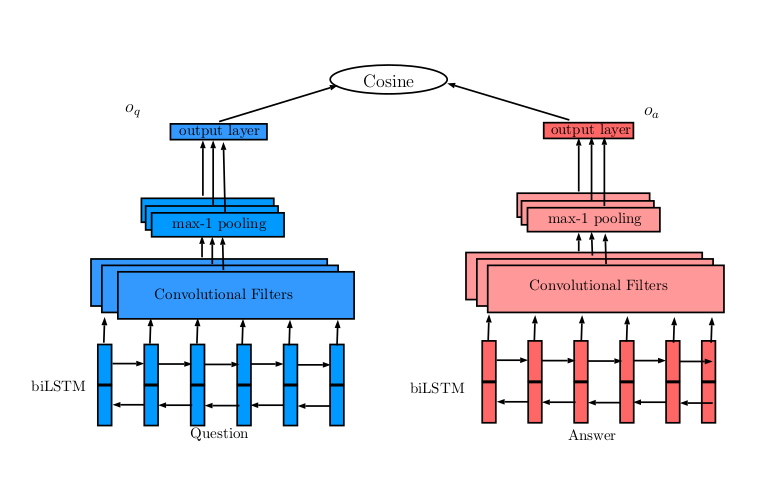
Pooling layers are known to suffer from an incapability to retain local linguistic information. To combat this the team proposed an additional CNN layer on top of the biLSTM layer. Additionally, they added a simple and efficient attention model to help their biLSTM model better distinguish between candidate answers according to the question context. This study is significant in the field of both deep learning based models and for the field of Question-Answering because it requires little or no feature engineering to achieve significant results. We will also see later in section [2.2](#X435a6cdd786300dff204ee7c2ef942d3e9034e2), how biLSTMs are implemented with Transformers to produce an even better model. In their work have shown that biLSTMs are capable of exploiting long-range sequential context information. This is important as it is often seen that the answer is not directly semantically related to the question but more contextually related. Long-range exploitations help the model understand the question more holistically, therefore helping produce a more relevant answer.

BiLSTMs have an advantage over traditional LSTMs in that they utilize both previous and future contextual information by processing the input in both directions, forward and backward. This leads to the generation of two independent sequences of output vectors form LSTMs. The final output is a concatenation of the outputs from both directions ie. . The **QA-LSTM** model proposed by can be seen in fig. [2.4](#lstmhaig). We can see that it generates a distributed representation for both questions(blue) and answers(red) before moving on to perform pooling and cosine similarity operations.



Basic QA-LSTM model depicting biLSTM implementation

In fig. [2.5](#lstmhaigcnn), it can be seen that the CNN layer has been implemented after the biLSTM layers and before the pooling layers. The pooling layers have been modified to be max- pooling layers, similar to standard CNNs.



QA-LSTM with CNN layer

The implementation of this convolution layer forces localized interactions between the inputs within a filter size of . For each window of size in biLSTM output vectors ie.

where is a time step and the convolution filter will generate one value as follows

Using -MaxPooling, the authors emphasized that a maximum of values will be kept for one filter, thereby indicating the highest degree that a filter matches the input sequence. To end this approach, there are parallel filters, with different parameter initialization which provide the CNN layer with -dimension output vectors. This produces two output vectors of dimension for each question and answer. Their experiments intuitively took as anything greater showed no improvements and it helped emphasize aspects of the answer such that this CNN based hybridbiLSTM could differentiate ground truths and incorrect answers efficiently. This

## Question Answering Using Transformers

Transformer based models have been become a very reliable way to implement various NLP based tasks since their introduction in the late 2017, . In this section we will critically evaluate a few key transformer based architectures & assess their application particularly on the SQuAD 2.0 dataset.

Work done in the field of Long short-term and gated recurrent and neural networks, in particular, has been established as a state of the art approach in sequence modelling, transduction problems such as language modelling and machine translation. Efforts have been made to push the boundaries of encoder-decoder based & recurrent language models.

Recurrent models usually adopt a combination approach where they take symbol position of the input and output positions along with the computation. This helps them align positions to steps in computation time and generate a sequence of hidden states , which is a function of the previous hidden state and the input function . This inherently sequential nature of RNNs causes memory issues, leading to reduced batch sizes.

Transformer Architecture built by

Transformer Architecture built by

The architecture for a *Transformer* in this paper is outlined as having an encoder that maps input sequences to a continuous representation. The architecture can be seen in Fig. [2.6](#transformerArchitecture). This is then decoded into an output sequence of symbols one at a time. Each step is auto-regressive, ie. it consumes the previously generated symbols as additional input when creating the next. This is similar to an ensemble model. Stacks of 6 encoder and 6 decoder layers is used.  
Each encoder layer has 2 sub-layers of a multi-head self-attention and the other a simple, position-wise fully connected feed-forward network layer. The output from each encoder layer can be represented as , where is the function implemented by the sub-layers in the model. Additionally, supporting residual connections is handled by producing outputs of dimension .

The decoder layer is similar to the encoder layer and has an additional 3rd sub-layer that performs multi-head attention over the output of the encoders. To ensure that the output layer isn’t affected by the output of residual connections from the sub-layers, a normalization layer is implemented at the end. Masking the subsequent positions and offsetting the output by 1 position ensures that the predictions for position can only depend on outputs from previous positions.

One of the key features in the Transformer architectures is *self-attention*. Let us look at this in detail below.

### Self Attention

The attention mechanism can be described as mapping a query to a set of key-value pairs. This can be seen from Fig. [2.7](#multiHeadAttention).  
The evaluations performed on the Wall Street Journal dataset, using 40k sentences, showed that even without task-specific tuning the model had better results with a fraction of the training cost.

Scale Dot and Multi-Head Attention Models

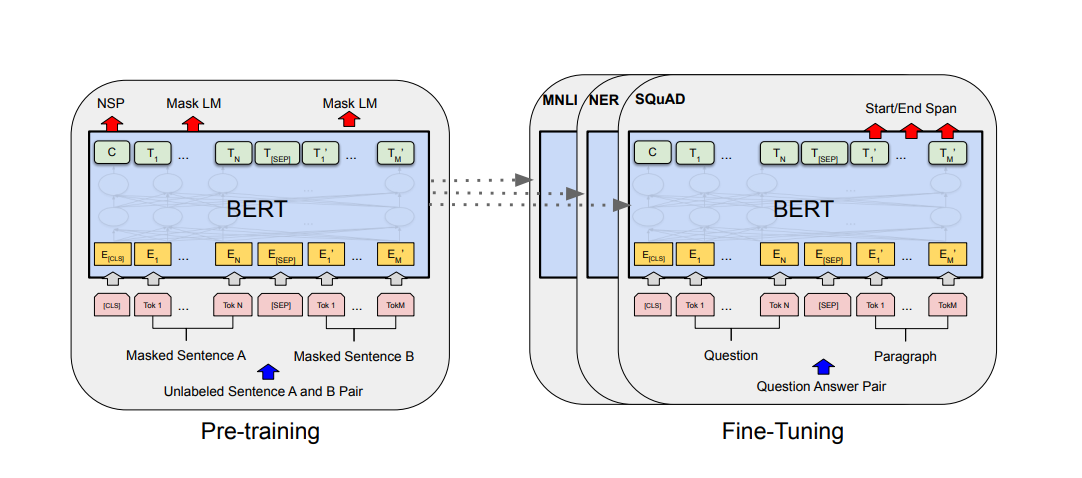
Scale Dot and Multi-Head Attention Models

### Improvements of Transformer Architectures

#### BERT

The paper on BERT, which is *Bidirectional Encoder Representations from Transformers* , introduces a new language model. This model is truly fascinating in many ways. First and foremost it is designed to pre-train deep bidirectional representations using unlabelled data. This is done by jointly conditioning context in all layers to the right and left. This pre-training allows the model to be fine-tuned simply using one additional output layer. These features make this model conceptually simple and very powerful empirically.

BERT employs language pre-training which has shown significant advantages in many applications e.g. paraphrasing, language level inference etc. These tasks aim to highlight the relationships between sentences through contextual understanding as well as by using tokenized outputs.



Pre-training and Fine Tuning procedures for BERT

BERT was tested on The General Language Understanding Evaluation (GLUE) benchmark which has a large number of diverse NLU tasks. BERT performed extremely well on the 11 NLP tasks that the authors ran it against. it showed an average accuracy improvement of 4.5% and 7% when compared to the previous state of the art models. Results of BERT and its significant gains make it one of the best candidate models for NLU tasks.

#### ALBERT

#### ROBERTA

#### DISTILBERT

## Comparison Of Techniques

## Summary

# Research Methodology

**DONT DO ANY ANALYSIS HERE**

## Data Selection

We have selected the Stanford Question Answering Dataset (SQuAD). This is as a reading comprehension dataset based on Wikipedia articles. It is based on questions posed by crowd-workers on a set of articles. The answer to every question is a segment of text or span, from the corresponding reading passage, or the question might be unanswerable .

The dataset consists of over 150,000 questions. Split into 100,000 answerable and 50,000+ unanswerable question, which were written to look similar to unanswerable questions. The challenge being that a model should be able to correctly answer the answerable questions and abstain from answering the unanswerable ones. The dataset is freely available as a part of the Transformers package in python or it can be downloaded from the SQuAD 2.0 website .

To effectively use this dataset for our purposes, let us first take a look at what its contents look like below.  
  
  
  
The answer to the aforementioned question is quite simple for humans to comprehend. The challenge is for us to contextualize this and make it machine-understandable so that our model can answer it correctly.

The dataset consists of various kinds of English language examples like negation, antonyms, entity swaps, impossible conditions to answer, answerable, etc. making the dataset a well-balanced one.

To use this dataset correctly we shall perform the following pre-processing steps on it:

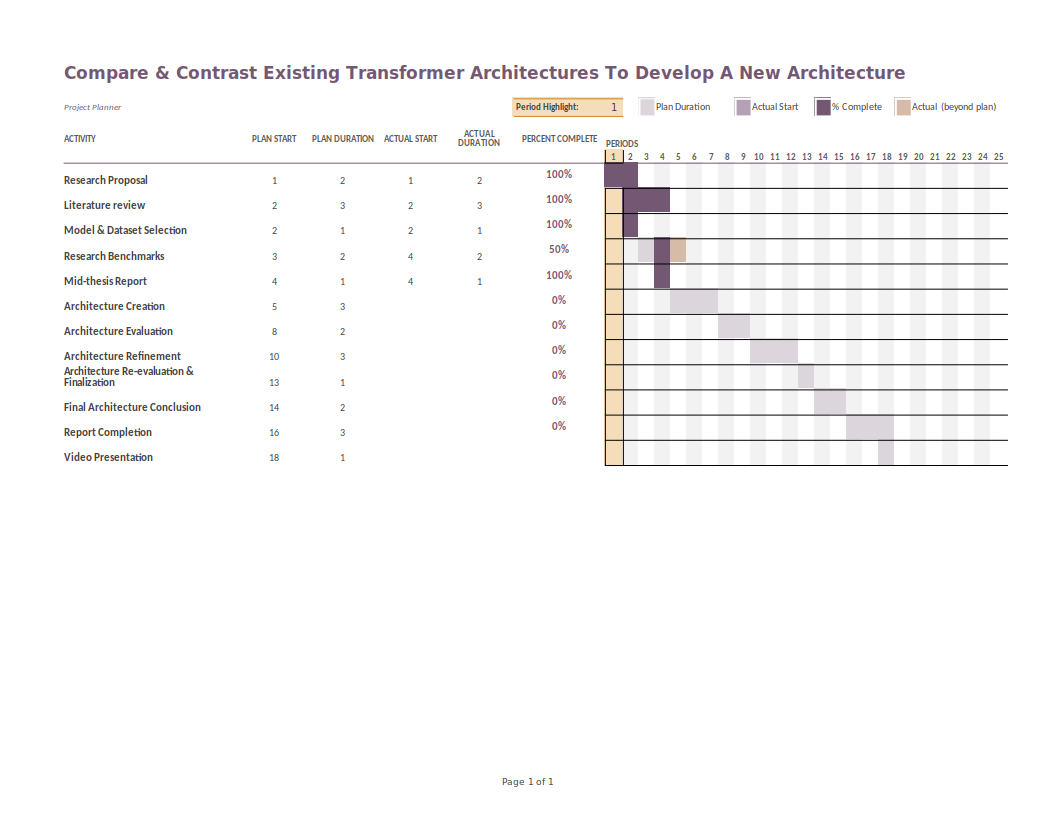
1. Data splitting into separate Question, Answer and Context lists.
2. Splitting the data into separate training and validation sets of question and answers using the 80/20 rule, also known as the Pareto principle. We will have 80% training data and 20% test data.
3. Tokenization of the split data to generate "context-question" pairs
4. Generating indexes for when an answer begins and ends in the dataset
5. Adding answer tokens based on their encoded positions

The SQuAD 2.0 Dataset , was developed with funding from Facebook to help address some major issues with existing datasets. Most datasets focus on questions that can be easily answered or use of automatically generated, unanswerable questions which are easily identifiable.  
The SQuAD 2.0 dataset resolves this by combining the SQuAD dataset along with 50,000 crowd worker generated unanswerable questions. The key feature of these being that the unanswerable questions must look similar to answerable ones. For a model to be successful on this new dataset, it must be able to answer all possibly answerable questions as well as determine when no answers are provided for a question in the given paragraph and abstain from answering. A comparative study was done for a Natural Language Understanding(NLU) task that obtained an 86% score on SQuAD 1.1, only got 66% on the new 2.0 dataset. The dataset helps bridge the gap between true NLU and machine understanding by using the concept of Relevance. Through comparisons with various datasets such as RACE, MCTest, QASENT etc. they have identified the missing links like negative examples, antonyms and helped fill the gap. This dataset forces the models to understand whether a paragraph span has the answer to the question posed.

## Data Pre-processing And Transformation

Techniques used. (1) Average pooling; (2) max pooling; (3) the concatenation of the last vectors on both directions. The three strategies are compared with the experimental performance in Section 5. Dropout operation is performed on the QA representations before cosine similarity matching

## Existing Models And Benchmarks



Mid-Thesis report Gantt Chart