NLP NATURAL LANGUAGE PROCESSING

CORE-GPT

GENERATIVELY PRE-TRAINED TRANSFORMER

Transformer based Character level Language Model

Submitted By :
Aditi Agrawal MCS22001
Kritika Tanwar MCS22016
Arundhati Warade MCS22021

WHY TRANSFORMER?

- Has shown to out perform than the **Traditional RNN** and **CNN** In may tasks.

- The key innovation of transformer is the **attention mechanism**, which allows the model to focus on different parts of the input sequence when making the predictions.

- It makes the transformer **well suited** for the tasks that requires the **long-range dependencies** (between the words in the sentences) such as text generation, machine translation, question and answering.

INTRODUCTION

WHAT IS TEXT SUMMARIZATION?

- Text summarization is a form of automatic summarization.
- Here, computer program processes a text.
- Generates a summary without any human intervention.

GOAL OF TEXT SUMMARIZATION

- The goal of text summarization is to identify the most important information in a document.
- Present it in a condensed form.
- Making the text easier for readers to comprehend the main points quickly.

CHALLENGES IN TEXT SUMMARIZATION

It requires understanding the context and meaning of a text and determining which information is important and relevant to the summary.

- Time-saving
- Information overload
- Easy comprehension
- Search engines
- Content curation

WHY TEXT SUMMARIZATION

APPLICATIONS

SEARCH ENGINES

To generate short summaries of web pages

NEWS AGGREGATION

To get a quick overview of the news without having to read the entire article.

BUSINESS INTELLIGENCE

Generate summaries that highlight the most important trends and insights.

LEGAL DOCUMENTS

To understand the essential details of a case or agreement.

SOCIAL MEDIA MONITORING

By capturing the main themes and sentiment of the conversation.

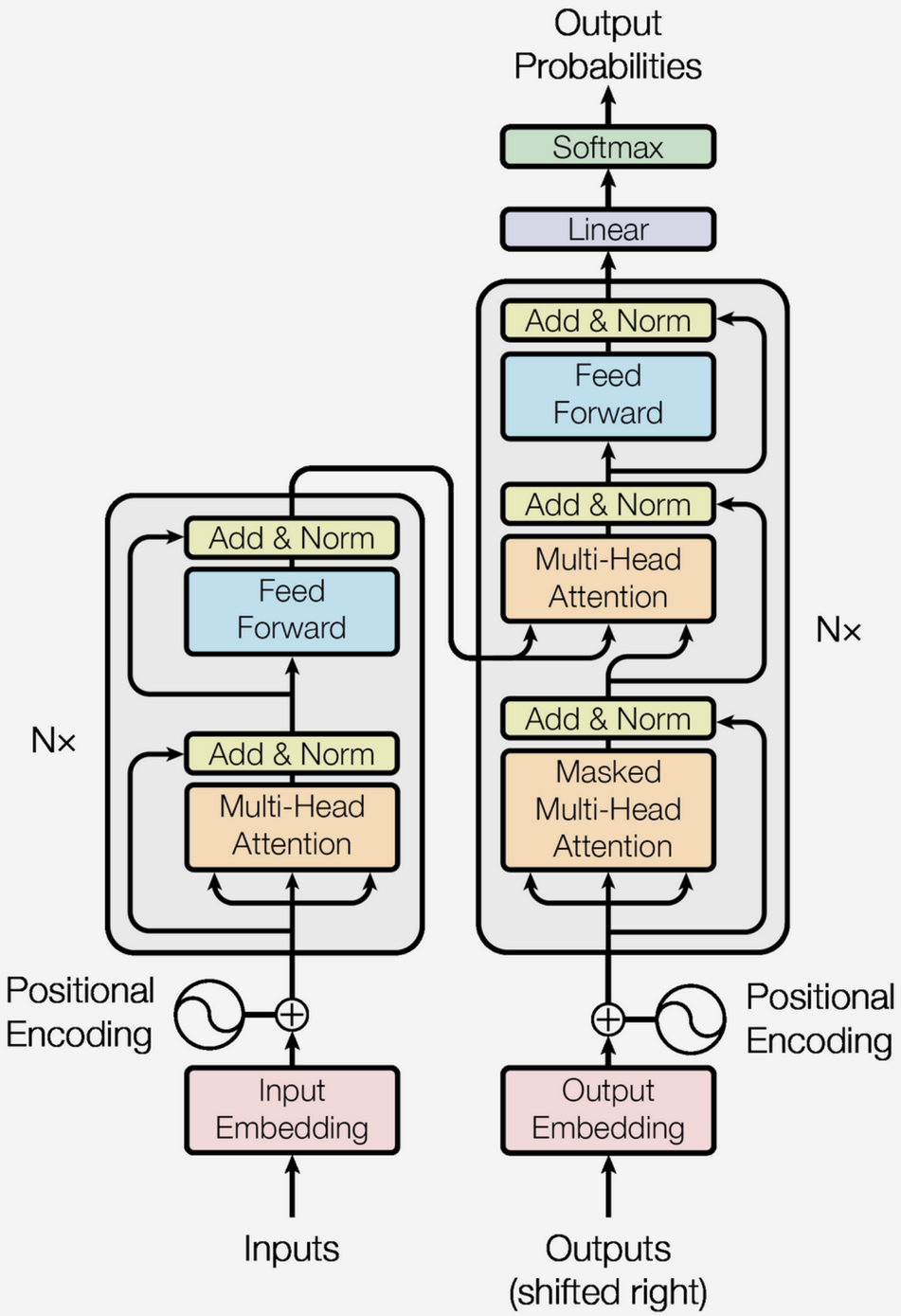
BIGRAM LANGUAGE MODEL

- A bigram language model is a statistical language model that predicts the probability of a word given the previous word in a sequence of words.

$$P(w_n|w_1^{n-1}) \quad \text{by} \quad P(w_n|w_{n-1})$$

- P (unicorn | the mythical) by P (unicorn | mythical)
- Markov assumption: the probability of a word depends only on the probability of a limited history
- Generalization: the probability of a word depends only on the probability of the n previous words
  - trigrams, 4-grams, ...
  - the higher n is, the more data needed to train
  - backoff models

TRANSFORMER MODEL



Transformers: unlike traditional models that process text data sequentially, transformers use a self-attention mechanism to weigh the importance of different parts of the input sequence.

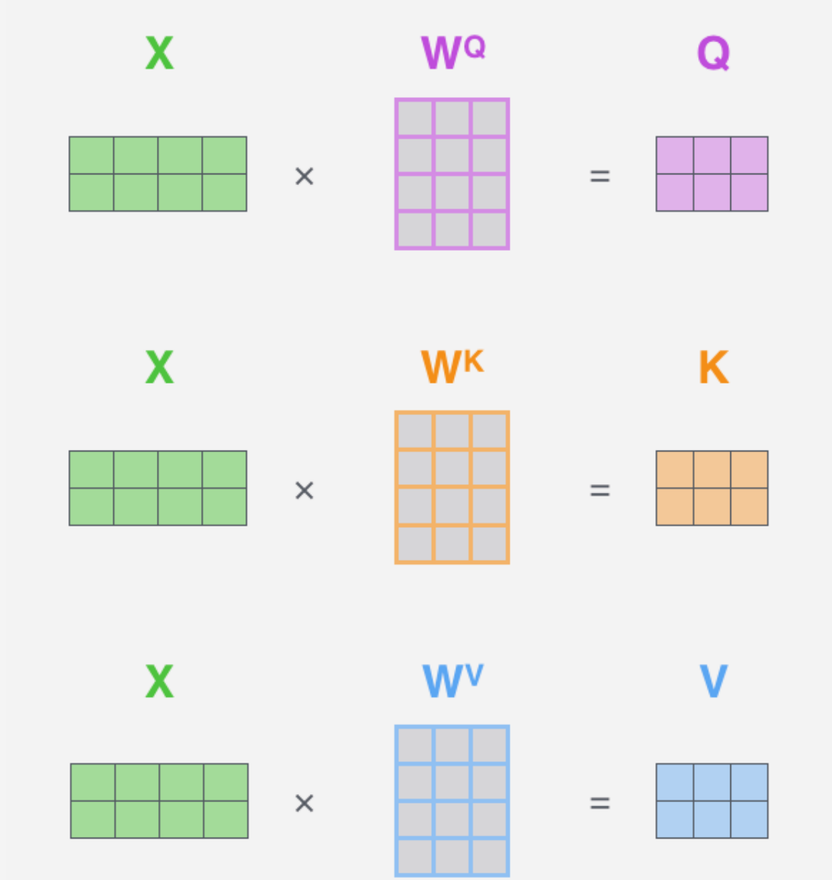
Here are some key features of transformers in NLP:

- Self-attention/Multi-head attention mechanism
- Encoder-decoder architecture
- Pre-training
- Fine-tuning
- State-of-the-art performance

ATTENTION AND THEIR TYPES

Self-attention:

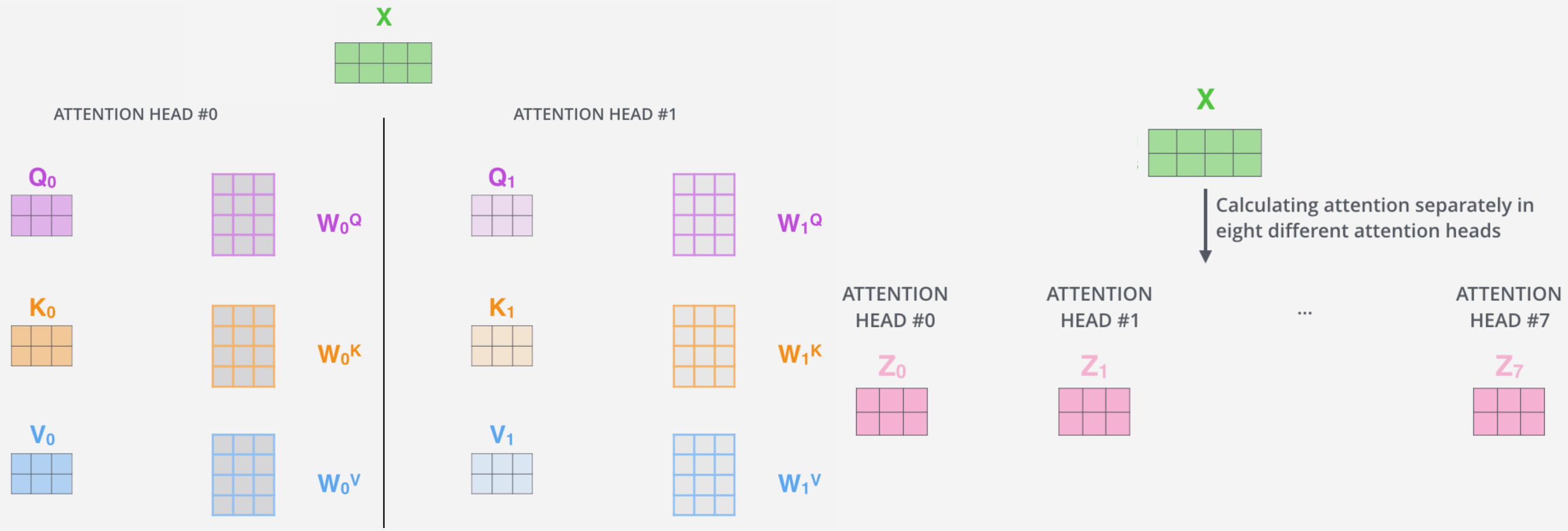
- It is a mechanism that allows the transformer to weigh the importance of different words in a sentence based on their relationships with other words in the sentence.
- It computes the attention score for each word in the sentence by comparing it with all other words in the sentence.



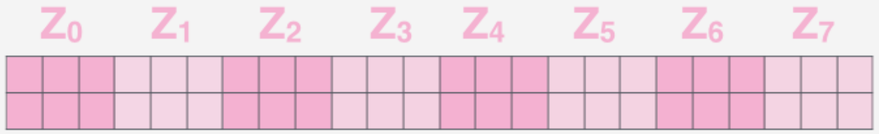
$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$

Multi-head attention:

- It is an extension of self-attention that allows the model to attend to information from different representation subspaces at different positions.
- It involves splitting the input representations into multiple heads, each of which attends to a different subspace of the input.

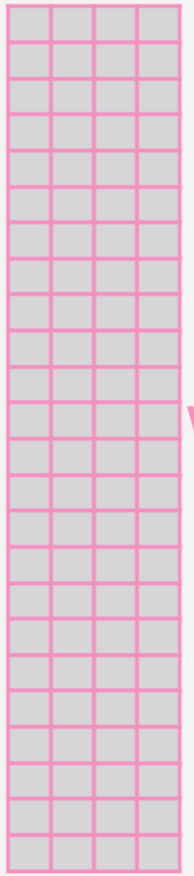


1) Concatenate all the attention heads

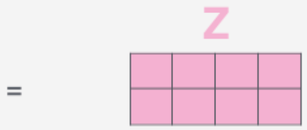


2) Multiply with a weight matrix W^O that was trained jointly with the model

X



3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN



Masked multi-head attention:

- It is a modification of multi-head attention that is used in the decoder layer of the transformer architecture.
- It involves masking out the attention scores for any positions that have not been generated yet during decoding.

Encoder-Decoder Attention:

- It works just like multiheaded self-attention, except it creates its Queries matrix from the layer below it, and takes the Keys and Values matrix from the output of the encoder stack.

MATHEMATICALLY

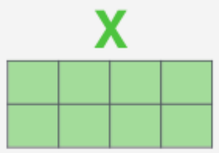
1) This is our input sentence*

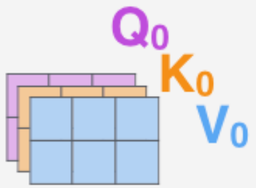
2) We embed each word*

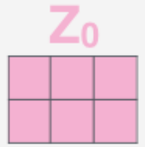
3) Split into 8 heads. We multiply X or R with weight matrices

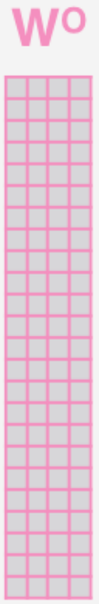
4) Calculate attention using the resulting Q/K/V matrices

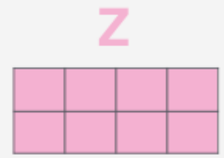
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

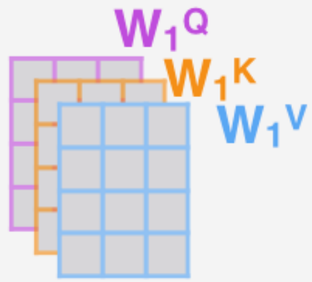


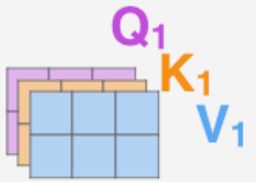


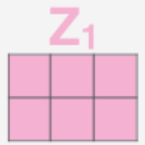




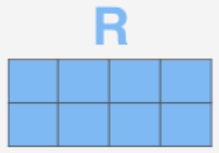


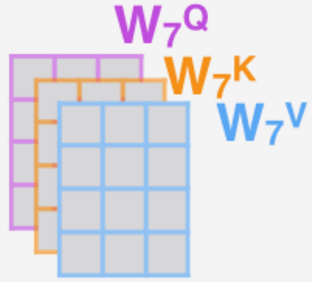


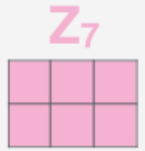




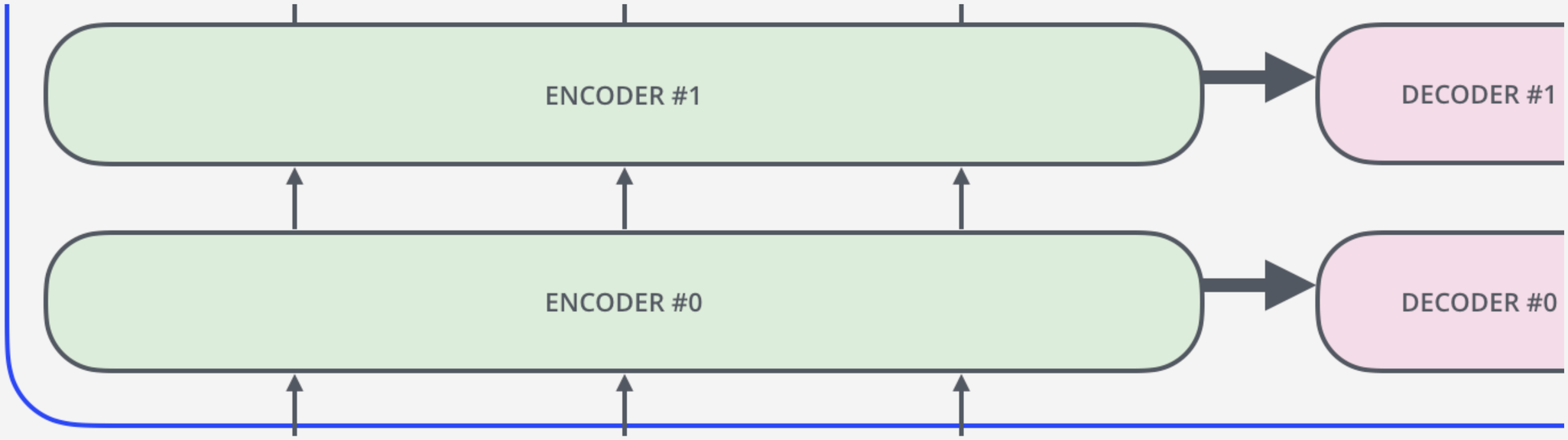
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

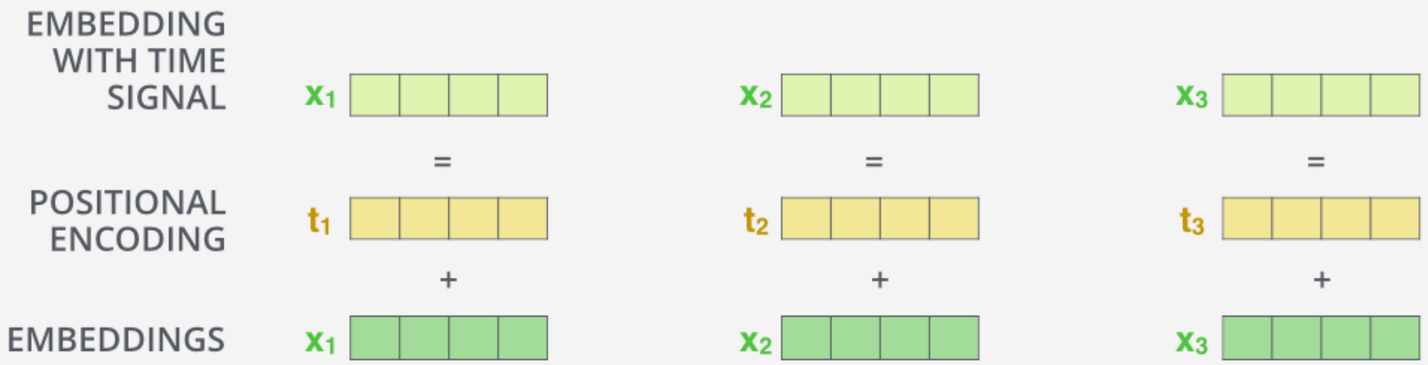




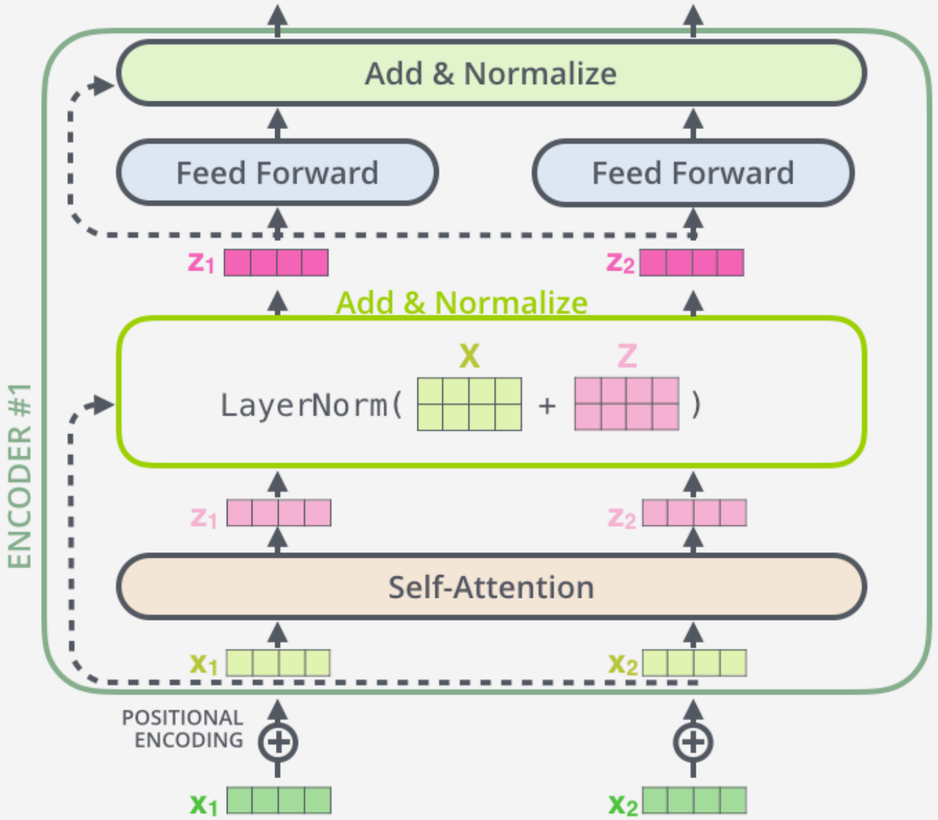


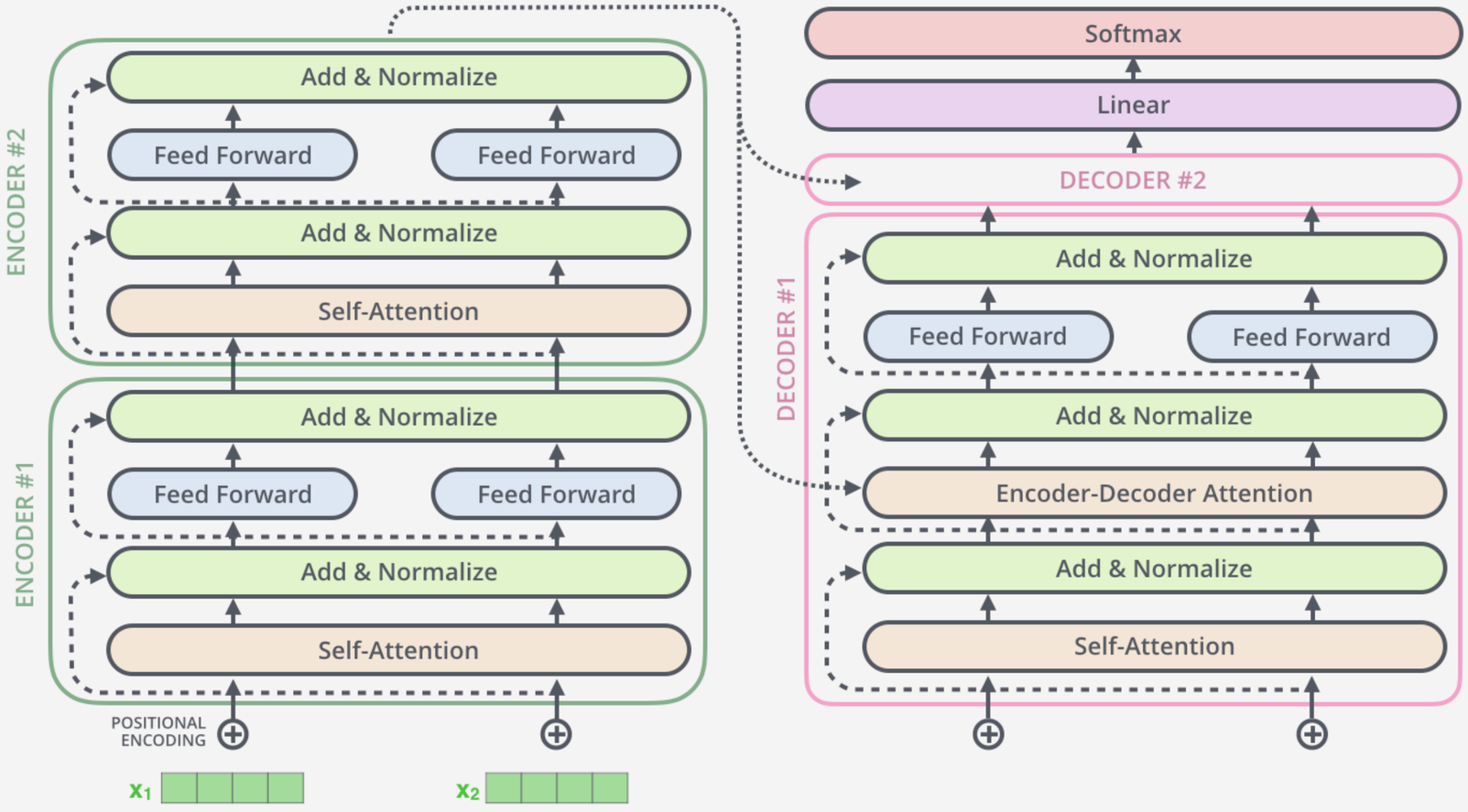
POSITIONAL ENCODING





THE RESIDUALS





DATASET

```
APTER ONE

e Boy Who Lived

    Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they
    Mr. Dursley was the director of a firm called Grunnings, which made drills. He w
    The Dursleys had everything they wanted, but they also had a secret, and their g
    When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, th
    None of them noticed a large, tawny owl flutter past the window.
    At half past eight, Mr. Dursley picked up his briefcase, pecked Mrs. Dursley on
  "Little tyke," chortled Mr. Dursley as he left the house. He got into his car and
    It was on the corner of the street that he noticed the first sign of something p
    But on the edge of town, drills were driven out of his mind by something else. A
    Mr. Dursley always sat with his back to the window in his office on the ninth fl
    He'd forgotten all about the people in cloaks until he passed a group of them ne
    "The Potters, that's right, that's what I heard —"
    " — yes, their son, Harry —"
    Mr. Dursley stopped dead. Fear flooded him. He looked back at the whisperers as
    He dashed back across the road, hurried up to his office, snapped at his secreta
upid. Potter wasn't such an unusual name. He was sure there were lots of people call
```

Input Data:

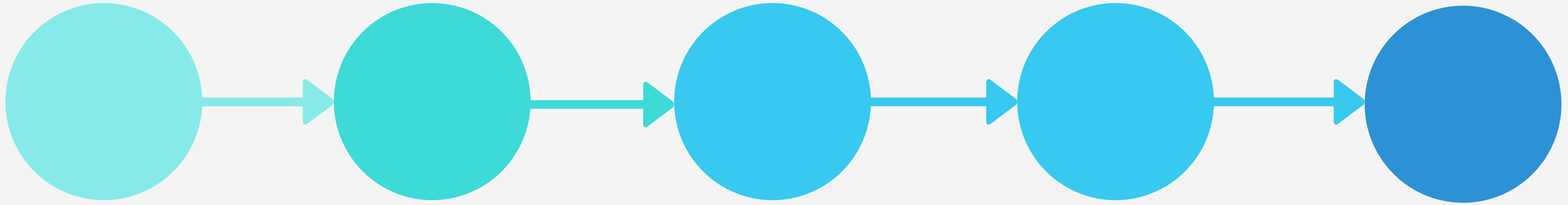
- Content: Harry Potter Part(1-4)
- No of Characters - 2712401
- No of Words: 479427
- Size of File: 2MB

HYPERPARMATERS

Batch_size	**number of paraller processing for chunks**	16
block_size	**size of chunks of data/ max word length possible here**	32
max_iters	**number of iterations to train the model**	**5000**
learning_rate	**governs the pace at which algorithm updates**	**1e-3**
n_embd	**total dimensions of the model**	**64**
n_head	**number of parallel attention heads**	**4**
n_layer	**normalized shape for LayerNorm**	**4**
dropout	**probability of an element to be zeroed**	**0.0**

METHODOLOGY

Bigram Model



Read a text file

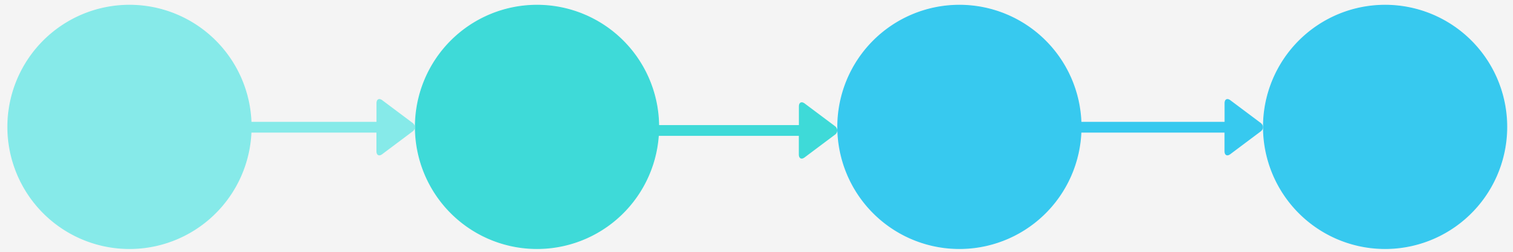
Tokenize the character

Generate input& target matrix

Position Encoding

Generate next character

Masked Self- Attention Module



Step 1,2,3 from above

Weighted Aggregation

Softmax

Generate Next Character using different models

METHODOLOGY

MODELS USED

Self Attention	The weights are generated from the inputs only.
Masked Multi Head Attention	Class that uses Normalization and Softmax
nn.Linear	Applies Linear Transformation to incoming data. ForMulti Head Attention
nn.Dropout	Randomly zeroes some values of input with probability p.
nn.LayerNorm	Mean and standard deviation are computed over the last 2 dimaensiosn of input
nn.Sequential	creates a Neural Network Model

OUTPUT

Some correctly identified words are:

- opened
- disk
- witch
- watching
- Snape
- Hermione
- Harry
- Dudley
- Malfoy
- Fred
- wizard

puning disk? "But Seoubled witch from withouth it Harry still, spottle very printen lamblef conted and bookshers at her watching excritched to Snape you'll got bit sell yBirsellivey slide of stop his importion. He wond try fillow is side it," said Hermione, skiday's ill, Tould it'll over guard peoping an well. "But — — no libjectle." "Didn't just not usued that his fack him as their what you was sorthing had moilimed nelem.

He shoulnds witch over his me," said Hellics supposed sight was almo the rothing sHortterta house, the doek in on for anyone in a was buddy didn't wantled and oncentack it've was be in, and got a students of when, at Havid pleipted Hermione was no wn Fred Sorcust heared.

"Oll now as prouse. It's nothing seat I know," shao. Harry was sent on the way and highed on it, severing again hopperets. "If you twernth tapped witchling extancious almisson osen very a nercle recond, "Gear. Afther it was perfer you was a af tratte pirruicolment speating some, exacting too in? I coult something his his hold happedings wit "You he such the chant crummes for tething. He would he looky mean spercied an thought Quitdituts putfonaging squestiens seemember screamerzed twill was plepped and thing ask they fingy begs2ad outsed of gasped. "Mo, win you bory you with o' Malfoy, to bit with, Harry, S1 Hercy and Great ditly wants on teres. Mave hair stoping out of Harry was to complesed bed wo his twell with is handfulorthing know they cage something into her fell, getting not broom, but it. How as thought, wildeger was stop for wording impromingstion," crapess Dudley backs me. "They be the be book day wisting into wiffe, sitty Lockhockher, Professor Snappeds peactent. pig? Harrid's for far wants commastly, but, think just Rod. There's stay wizardly armpm and Malfoy cleaped a lirge?" Harry rememednong to get the tick aroub her threw, which wizard im. A wellinco halm will inly end."

But of twer been the ctut shand lasmed the cloar and him forn aw hack time tacks. Besting wond you held said, He one didnea ed," said Hermionione whispers. "Wellibbicklely.

He sokin's got of Fon, back to well. Que it fall a scaulds verice hears

RESULTS

Model	Train Loss	Validation Loss
Bigram	2.4738	2.4911
Self Attention	2.3980	2.4084
Multi Headed Atttention	2.2748	2.2858
Self Atention + Feed Forward	2.2286	2.22412
Multi Head + Feed Forward	1.9993	2.0808
Layer Norm + Feed Forward + Multi Head	1.9827	2.0607
Scaling Transformers	1.4963	1.5129

Thank You