

COMPARATIVE STUDY BETWEEN BERT AND ROBERTA MODELS FOR QUESTION ANSWERING

Kishore Kumar Kalathur Chenchu
kumar.278@iitj.ac.in

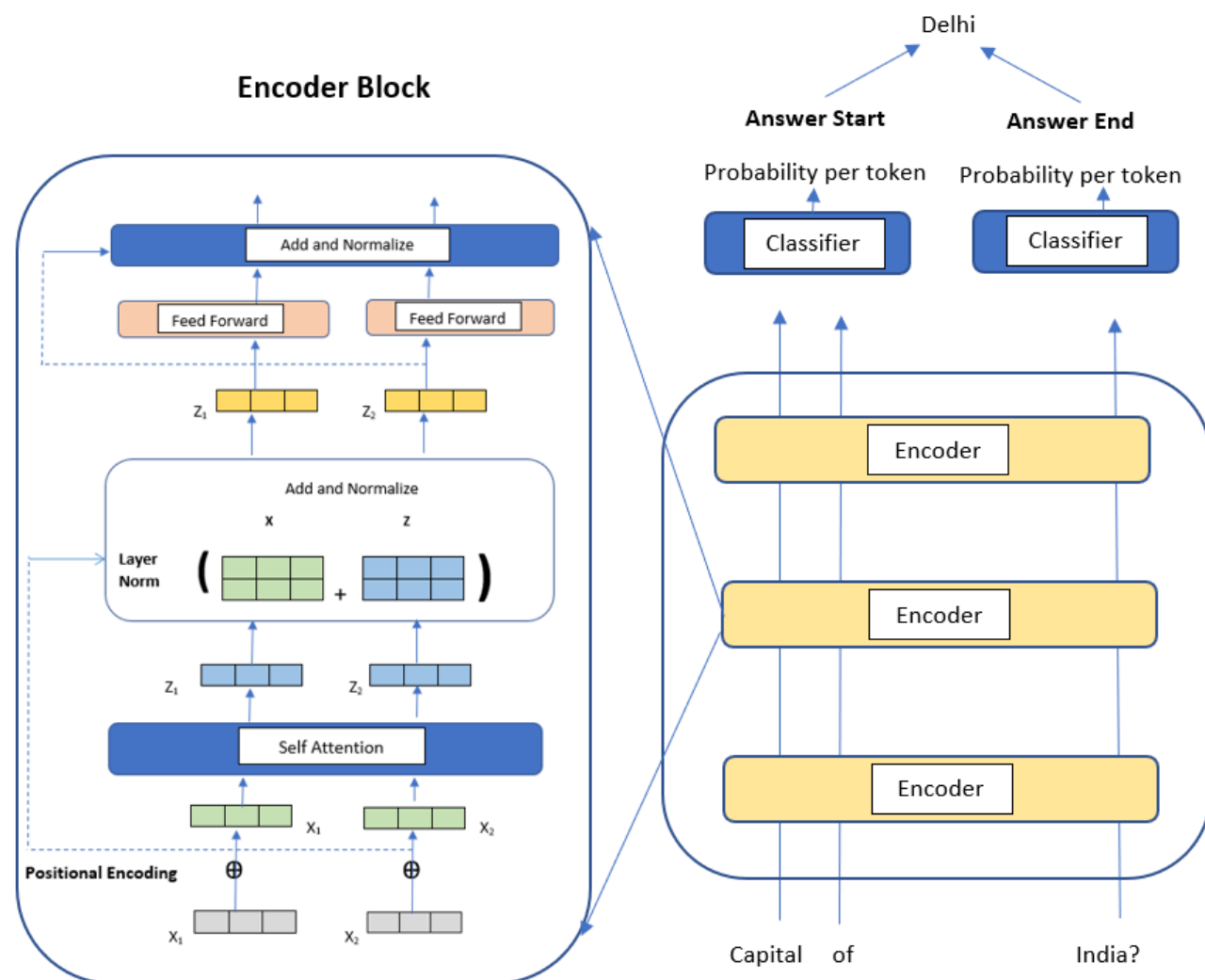
Prabhala Sandhya Gayatri
gayatri.1@iitj.ac.in

Tejaswee A
tejaswee.1@iitj.ac.in

Objective

Objective of this project is to answer the questions asked to a machine using BERT or RoBERTa approaches.
Input - Questions in English or arithmetic statements
Output - Answers to the asked questions

Architecture Used



BERT vs RoBERTa

BERTBASE (L =12, H = 768, A = 12, 110M parameters)

BERT implementation performs a single static mask. Each training sequence is seen with the same mask four times during training.

ROBERTA (L = 24,H = 1024, A = 16, 355M parameters)

RoBERTa is trained with dynamic masking by generating the mask pattern corresponding to every input sequence that we feed the model without NSP loss, large mini-batches and a larger byte-level BPE

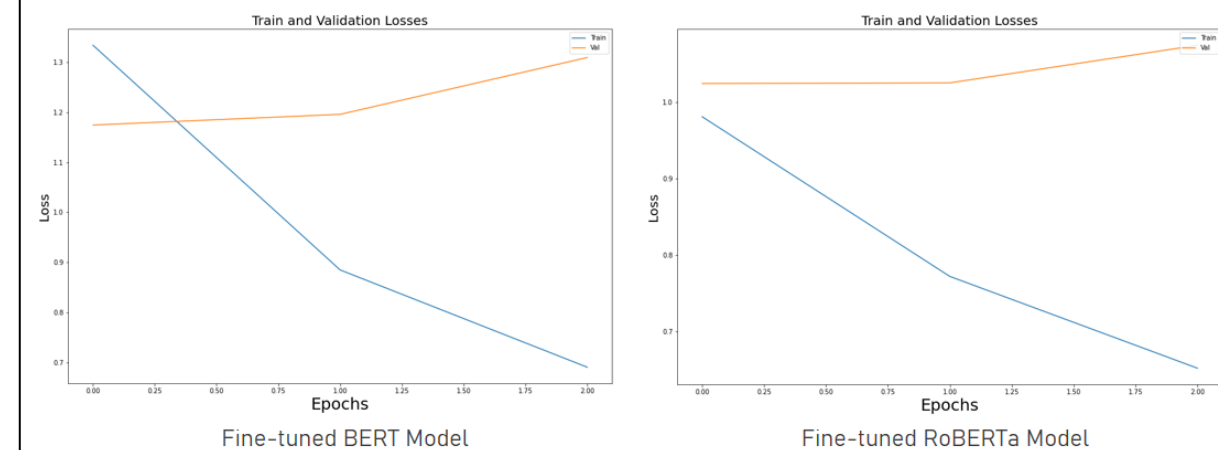
Experiments

	Train Loss	Validation Loss
Epoch 1		
BERT	1.33	1.174
RoBERTa	0.98	1.024

	Train Loss	Validation Loss
Epoch 2		
BERT	0.88	1.19
RoBERTa	0.77	1.02

	Train Loss	Validation Loss
Epoch 3		
BERT	0.6905	1.309
RoBERTa	0.65	1.07

Train and Validation Loss Plots



— Train Loss
— Validation Loss