EasyChair Preprint

№ 9323

GPT-K : a GPT Based Model for Generation of Text in Kannada

K H Manodnya and Animesh Giri

November 15, 2022

GPT-K: A GPT-based model for generation of text in Kannada

Manodnya K H
Department of Computer Science and Engineering
PES University
Bangalore, India
manodynak@gmail.com

Animesh Giri
Department of Computer Science and Engineering
PES University
Bangalore, India
animeshgiri@pes.edu

***Abstract*— Large AI-based language models are changing how we work with language. They are becoming increasingly popular because they allow us to create complex linguistic structures without requiring a lot of resources. A language model must have access to a large corpus of linguistic data (e.g., word frequencies) to learn and generate new words. GPT-2, a language model, can generate coherent paragraphs independently, without any input on what to write about or guidance on grammar rules. Although multiple pre-trained GPT-2 models exist for English and other high-resource languages, there are few to no such models for Indic languages like Kannada. In this study, we propose GPT-K, a GPT-2 based model for language modeling in Kannada. GPT-K has been trained on a large corpus of Kannada text and can effectively perform language modeling tasks in Kannada. The model generated syntactically correct text in most cases**.

Keywords—GPT-2, large language models, language modeling, model training, hyperparameter finetuning, Indic languages

I. INTRODUCTION

Large language models are transforming the way we think about language. They can do things like generate text or transform text. This is changing how we work with language and has garnered a lot of attention and become increasingly popular in the past few years. Large Language Models use machine learning algorithms to process enormous text-based sets. These models can understand, predict, and generate human languages by processing a massive text corpus. They are increasingly based on transformer-based architectures, which can deal with the excessive amount of text that such models require. They must have access to a large corpus of linguistic data (e.g., word frequencies) to learn and generate new words.

Language models are generated from text corpora and trained either with supervised learning algorithms like maximum entropy or conditional random fields or through unsupervised training. A transformer is an algorithm that takes in a sequence of words and outputs another sequence of words. The GPT-2, created by OpenAI, is an example of a transformer. It can generate coherent paragraphs and even short stories on its own, without any input on what to write about or guidance on grammar rules. GPT-2 can generate coherent paragraphs of natural-sounding text in any language with 99% accuracy after just 10 minutes of reading training data.

When GPT-2 is fed with an input sentence, it takes the words and rearranges them to generate a new sequence of sentences. Although GPT-2 was initially designed for translation purposes, it can be finetuned for use in other applications as well: summarization, paraphrasing, and text generation. GPT-2 has been shown to be able to generate coherent and grammatically correct paragraphs of English text, which are comparable in quality to those generated by human copywriters.

Although multiple pre-trained GPT-2 models exist for English and other high-resource languages, there are few to no such models for Indic languages like Kannada. In this study, we propose GPT-K, a GPT-2 based model for language modeling in Kannada. GPT-K has been trained on a large corpus of Kannada text and can effectively perform language modeling tasks in Kannada.

The major research contributions of this study are summarized as follows:

1. Collation of text-based datasets in the Kannada language to generate a large corpus of Kannada text for training.
2. Preprocessing of datasets to eliminate unwanted text and invalid characters.
3. Finetuning the hyperparameters for optimal performance.
4. Finetuning the GPT-2 model to reduce compute costs.
5. Training the finetuned GPT-2 model in the Kannada language.
6. Evaluating the model.

II. RELATED WORKS

Vaswani et al., 2017 [1] proposed a new simple network architecture solely based on attention mechanisms doing away with recurrence and convolution entirely. Experiments showed that these models were superior in quality and required less training time. They were also more parallelizable. They

achieved a BLEU score of 41.0, surpassing the best models of the time.

Radford et al., 2018 [2], in their paper "Improving Language Understanding by Generative Pre-Training," demonstrated that significant gains on tasks like document classification, question answering, textual entailment, and semantic similarity assessment can be achieved by generative pre-training on a large corpus of unlabeled text followed by discriminative fine-tuning on each specific task. The proposed model was named GPT.

Radford et al., 2019 [3] proposed a new language model GPT-2 and also demonstrated that language models begin to learn Natural language processing tasks such as reading comprehension, machine translation, question answering, and summarization without any supervision. Their largest model, GPT-2, is a 1.5B parameter Transformer that achieves astonishing results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText, an internal OpenAI corpus. To create this corpus, all outbound links from Reddit with karma greater than 3 were scraped. Samples from the model contain articulate paragraphs of text and reflect these improvements; Their findings suggested a promising path toward building language processing systems which, from their naturally occurring environment, learn to perform tasks. Although the largest model had 1.5 billion parameters, the largest model open-sourced by OpenAI had only 774 million parameters. Concerns over the potential misuse of the technology were cited for not releasing the larger models.

Brown et al., 2020 [4] proposed GPT-3, which is architecturally similar to GPT-2 except that Alternating dense and locally banded sparse attention patterns were used in GPT-3. This autoregressive language model was trained on 175 billion parameters and outperformed GPT-2 on most counts. However, this model was not open-sourced due to its potential for misuse, and exclusive rights to use the model were granted to Microsoft corporation.

So et al., in their 2022 paper "Primer: Searching for Efficient Transformers for Language Modeling" [5], proposed a new language model called Primer(PRIMitives searched transformER) in which they demonstrated that by squaring ReLU(Regularized evoLUtion) activations and adding a depthwise convolution layer after each Q, K, and V projection in self-attention, the training cost for transformers could be significantly reduced.

Liao et al., 2019 [6] propose a GPT-based generation for classical Chinese poetry. They use a simple GPT model [2] to generate various forms of classical Chinese poems that meet form and content requirements. While retaining the GPT architecture, they only fine-tune the model on a large corpus of Chinese poetry.

Dhivyaa et al., 2022 [7] propose an attention-based LSTM-NMT model for Tamil text summarization based on the GPT-2 architecture. They propose an improved GPT-2 model to perform text summarization. They propose an efficient model for text summarization in an Indic language, Tamil. Their model uses an attention-based LSTM-NMT model for transliterating Tamil text to English text which is then processed by GPT-2 and later translated back to Tamil. They improve the existing GPT-2 architecture by adding a masked self-attention layer to the decoder block, allowing for large batch sizes and parallel processing of multiple tokens.

III. Our Method

The study can be broadly divided into six phases, model selection, data collection and preprocessing model finetuning, training, hyperparameter finetuning, and model evaluation. We use an improved GPT-2 model for this study. We call this model GPT-K. This model is then trained on a large corpus of Kannada text to generate text samples in Kannada. We adopt GPT-2's vocabulary and tokenization.

A. Model details

GPT-2 [3], based on the transformer architecture[1], is the basic model used for our study. Figure 1 shows the basic model proposed in [3]. Equation 1 is used by the authors of [3] to factorize the joint probabilities over symbols as the product of conditional probabilities since, language has a natural sequential ordering. In the equation, x represents the set of variables, whereas S represents the set of samples. We use the open-sourced 117Million parameters GPT-2 model for this study. Depth-wise convolution layers are then added after each K, Q, and V projections in self-attention as proposed in [5]. ReLU(Regularized evolution) activations are also squared as proposed in [5]. These changes reduce compute requirements significantly. We adopt the tokenization and encoding for UTF-8-based character sets released in GPT-2 for encoding Kannada text. Figure 2 shows the improved GPT architecture.

$$p(x) = \prod_{i=1}^{n} p(S_n|S_1, \dots, S_{n-1})$$

Equation 1.

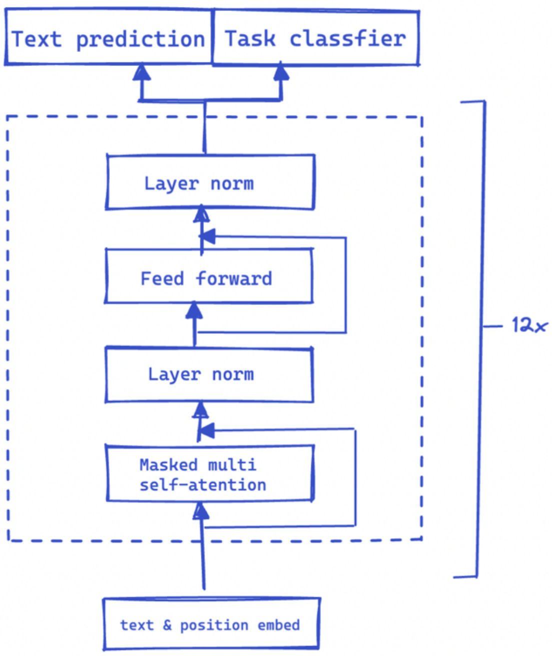

Figure. 1 GPT-2 architecture

B. Data collection and preprocessing

Datasets available on the web, OSCAR corpus [8], CC-100 [9], and the Kannada Wikipedia dump [10] were collected. These datasets were then cleaned and preprocessed. All the metadata was stripped while retaining only Kannada text which was then written to a set of text files, rendering a continuous corpus of cleaned text. This data was preprocessed, and byte-pair encoding (BPE) was used to encode these files. Invalid UTF-8 characters were ignored in this process, and only the valid UTF-8 bits were encoded. The encoded text was then written to a single file. Figures 3 and 4 show samples of the raw and cleaned corpus, respectively.

{"label":"kn","prob":0.99815625},{"label":"kn","prob":0.9998685},{"label":"kn","prob":0.99964076},
{"label":"kn","prob":0.9973867},{"label":"kn","prob":0.9984214},{"label":"kn","prob":1.0000386},
{"label":"kn","prob":0.9991355},{"label":"kn","prob":0.99988204},{"label":"kn","prob":0.999739},
{"label":"kn","prob":0.99834085},{"label":"kn","prob":0.99996656},{"label":"kn","prob":0.99939144},
{"label":"kn","prob":0.99922496},{"label":"kn","prob":0.9995925},{"label":"kn","prob":0.9996038},
{"label":"kn","prob":0.9994538},{"label":"kn","prob":0.9999307},null,null,null,
{"label":"en","prob":0.90416515},null,null,null,
{"label":"en","prob":0.8193673},null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null,null]}}
{"content":"ಅದು ಕೇವಲ ನಾಲ್ಕು ನಿಮಿಷ ಐವತ್ತೊಂಬತ್ತು ಸೆಕೆಂಡು ನಡೆದ ಘಟನೆ. ಕಣ್ಮುಚ್ಚಿ ತೆರೆಯುವಷ್ಟರಲ್ಲಿ ನಡೆದ ಈ ಘಟನೆ ಮಂಗಳೂರಿನ ಕೋಮುವಾದಿಗಳ ಬಣ್ಣವನ್ನು ಬಯಲು ಮಾಡಿತ್ತು. ಆದರೆ ರಾಜ್ಯದ ಆಡಳಿತ ವ್ಯವಸ್ಥೆ ಈ ಕೋಮುವಾದಿಗಳ ಜೊತೆ ಸೇರಿಕೊಂಡು ಮಾಡುವ ಕಿರಾತಕ ಕಾರ್ಯಚಟುವಟಿಕೆಗಳನ್ನು ಈ ಘಟನೆ ಬಯಲಿಗೆ ತರಲೇ ಇಲ್ಲ. ಪೊಲೀಸರು ಮಾಧ್ಯಮದ ಪ್ರತಿನಿಧಿಗಳ ಮೇಲೆ ಕೇಸು ದಾಖಲಿಸಿ ಪ್ರಕರಣವನ್ನು ವಿಷಯಾಂತರ ಮಾಡಲು ಯತ್ನಿಸಿದ್ದರು. ಮಂಗಳೂರಿನ ಕೆಲವೊಂದು ಚೆಡ್ಡಿ ಪತ್ರಕರ್ತರ ಕೃಪೆಯಿಂದ ಅದರಲ್ಲಿ ಸ್ವಲ್ಪಮಟ್ಟಿಗೆ ಯಶಸ್ವಿಯೂ ಆದರು. ಆದರೆ ಸತ್ಯ ನಮ್ಮ ಕ್ಯಾಮರಾಗಳಲ್ಲಿ ಸೆರೆಯಾಗಿದೆ. ದಿನಾಂಕ, ಸಮಯ ಸಹಿತ ಪೊಲೀಸರೂ ಈ ಪ್ರಕರಣದಲ್ಲಿ ಭಾಗಿಯಾದ ಬಗ್ಗೆ ಕ್ಯಾಮರ ಕ್ಯಾಸೆಟ್ಟು ಸಾಕ್ಷಿ ನುಡಿಯುತ್ತದೆ.\nಮಂಗಳೂರಿನ ಮಾರ್ನಿಂಗ್ ಮಿಸ್ಟ್ ಹೋಂ ಸ್ಟೇಯಲ್ಲಿ ಹಿಂದೂ ಯುವತಿಯರ ಜೊತೆ ಮುಸ್ಲಿಂ ಯುವಕರು ಪಾರ್ಟಿ ಮಾಡುತ್ತಿದ್ದಾರೆ ಎಂಬ ಮಾಹಿತಿಯ ಮೇಲೆ ಹಿಂದೂ ಜಾರಣ ವೇದಿಕೆಯ ಕಾರ್ಯಕರ್ತರು ಮಾರ್ನಿಂಗ್ ಮಿಸ್ಟ್ ಪ್ರವೇಶ ಮಾಡಿದ್ದರು.}

Figure 3. Sample of the raw corpus

ಕನ್ನಡದ 'ಪ್ರಚಂಡ' ಕುಳ್ಳ ದ್ವಾರಕೀಶ್ ಮತ್ತೆ ನಿರ್ಮಾಣಕ್ಕೆ ಸಜ್ಜಾಗಿದ್ದಾರೆ. ವಿಷ್ಣುವರ್ಧನ್, ಸೌಂದರ್ಯಾ 'ಆಪ್ತಮಿತ್ರ' ಚಿತ್ರದ ಯಶಸ್ಸಿನ ನಂತರ ಆರೇಳು ವರ್ಷ ಸುಮ್ಮನಿದ್ದ ದ್ವಾರಕೀಶ್ ಕಳೆದ ವರ್ಷ, 2011 ರಲ್ಲಿ ಮತ್ತೆ ನಿರ್ಮಾಣಕ್ಕಿಳಿದು 'ಓನ್ಲಿ ವಿಷ್ಣುವರ್ಧನ' ಚಿತ್ರ ನಿರ್ಮಿಸಿದ್ದರು. ಆ ಚಿತ್ರವೀಗ ಶತಕ ಬಾರಿಸಿ ನಿರ್ಮಾಪಕ ದ್ವಾರ್ಕಿ ಮುಖದಲ್ಲಿ ಮಂದಹಾಸ ಮೂಡಲು ಕಾರಣವಾಗಿದೆ.
ಇದೀಗ ವಿಷ್ಣುವರ್ಧನದಲ್ಲಿ ನಾಯಕಿಯರಲ್ಲೊಬ್ಬರಾಗಿದ್ದ ಪ್ರಿಯಾಮಣಿಯನ್ನು ಪ್ರಧಾನವಾಗಿಟ್ಟುಕೊಂಡು 'ಚಾರುಲತಾ' ಎಂಬ ಸಿನಿಮಾ ನಿರ್ಮಾಣಕ್ಕೆ ಮುಂದಾಗಿರುವ ದ್ವಾರಕೀಶ್, ಇದೇ ತಿಂಗಳು ಅಂದರೆ ಮಾರ್ಚ್ 23, 2012 ರ ಯುಗಾದಿ ಶುಭದಿನದಂದು ಮುಹೂರ್ತವಿಟ್ಟುಕೊಂಡಿದ್ದಾರೆ. ನಾಯಕಿ ಪ್ರಧಾನವಾಗಿರುವುದು ಮಾತ್ರವಲ್ಲ, ನಾಯಕ ಇರುವುದೇ ಇಲ್ಲ ಎಂಬ ಸುದ್ದಿಯೂ ಹರಿದಾಡುತ್ತಿರುವುದು ಸುಳ್ಳಲ್ಲ.
ವಿಷ್ಣುವರ್ಧನ ಚಿತ್ರದ ನಿರ್ದೇಶಕ ಕುಮಾರ್ ಮೇಲೆ ಮತ್ತೊಮ್ಮೆ ಭರವಸೆಯಿಟ್ಟಿರುವ ದ್ವಾರ್ಕಿ, ಚಾರುಲತಾರನ್ನೂ ಕೂಡ ಅವರ ಕೈಗೆ ಒಪ್ಪಿಸುತ್ತಿದ್ದಾರೆ.
"ಇದೊಂದು ಪಕ್ಕಾ ಮನರಂಜನೆಯ ಚಿತ್ರ. ಪ್ರಿಯಾಮಣಿ ಈ ಹಿಂದೆ ನಟಿಸಿರದಂತ ಅಪರೂಪಕ್ಕೆ ಮಾತ್ರ ಸಿಗುವಂತೆ ಪಾತ್ರ" ಎಂದು ಹೇಳುವ ಮೂಲಕ ದ್ವಾರಕೀಶ್ ಪ್ರೇಕ್ಷಕರಲ್ಲಿ ಕುತೂಹಲದ ಜೊತೆ ನಿರೀಕ್ಷೆಯನ್ನೂ ಹುಟ್ಟುಹಾಕಿದ್ದಾರೆ
ಚಾರುಲತಾ ಚಿತ್ರಕ್ಕೆ ಹೆಸರಿಗೆ ಮಾತ್ರ ದ್ವಾರಕೀಶ್ ನಿರ್ಮಾಪಕ. ಆದರೆ ಕೆಲಸವನ್ನೆಲ್ಲ ನೋಡಿಕೊಳ್ಳಲಿರುವವರು ಕಾರ್ಯಕಾರಿ ನಿರ್ಮಾಪಕ ಎಂಬ ಹೆಸರಿನಲ್ಲಿ ಅವರ ಪುತ್ರ ಯೋಗೀಶ್. ತಮಿಳಿನ 'ನಾಡೋಡಿಗಳ್' ಚಿತ್ರಕ್ಕೆ ಸಂಗೀತ ನೀಡಿದ್ದ ಸುಂದರ್ ಸಿ. ಬಾಬು ಈ ಚಿತ್ರಕ್ಕೂ ಸಂಗೀತ ನಿರ್ದೇಶನ ಮಾಡಲಿದ್ದಾರೆ. ರಾಜರತ್ನಂ ಕ್ಯಾಮರಾ ಹಿಡಿಯಲಿದ್ದಾರೆ.
ಕಥೆ, ಚಿತ್ರಕಥೆಯ ಜತೆ ನಿರ್ದೇಶನದ ಹೊಣೆಯೂ ವಿ. ಕುಮಾರ್ ಅವರದ್ದು. ಇದೇ ಮಾರ್ಚ್ 23ರ ಯುಗಾದಿಯ ಶುಭ ದಿನದಂದು ಸೆಟ್ಟೇರುತ್ತಿರುವ ಚಾರುಲತಾಗೆ ಉಳಿದ ಪಾತ್ರಗಳ ಆಯ್ಕೆ ಸೇರಿದಂತೆ ಇನ್ನಿತರ ಸಂಗತಿಗಳನ್ನು ದ್ವಾರಕೀಶ್ ಶೀಘ್ರದಲ್ಲೇ ಬಹಿರಂಗಪಡಿಸಲಿದ್ದಾರಂತೆ. ಸದ್ಯಕ್ಕೆ ತಮ್ಮ ನಿರ್ಮಾಣದ ಕಿಚ್ಚ ಸುದೀಪ್ ನಾಯಕತ್ವದ 'ವಿಷ್ಣುವರ್ಧನ' ಶತಕ ಭಾರಿಸಿದ ಸಂತೋಷವನ್ನು ದ್ವಾರ್ಕಿ ಅನುಭವಿಸುತ್ತಿದ್ದಾರೆ.

Figure 4. Sample from the cleaned corpus

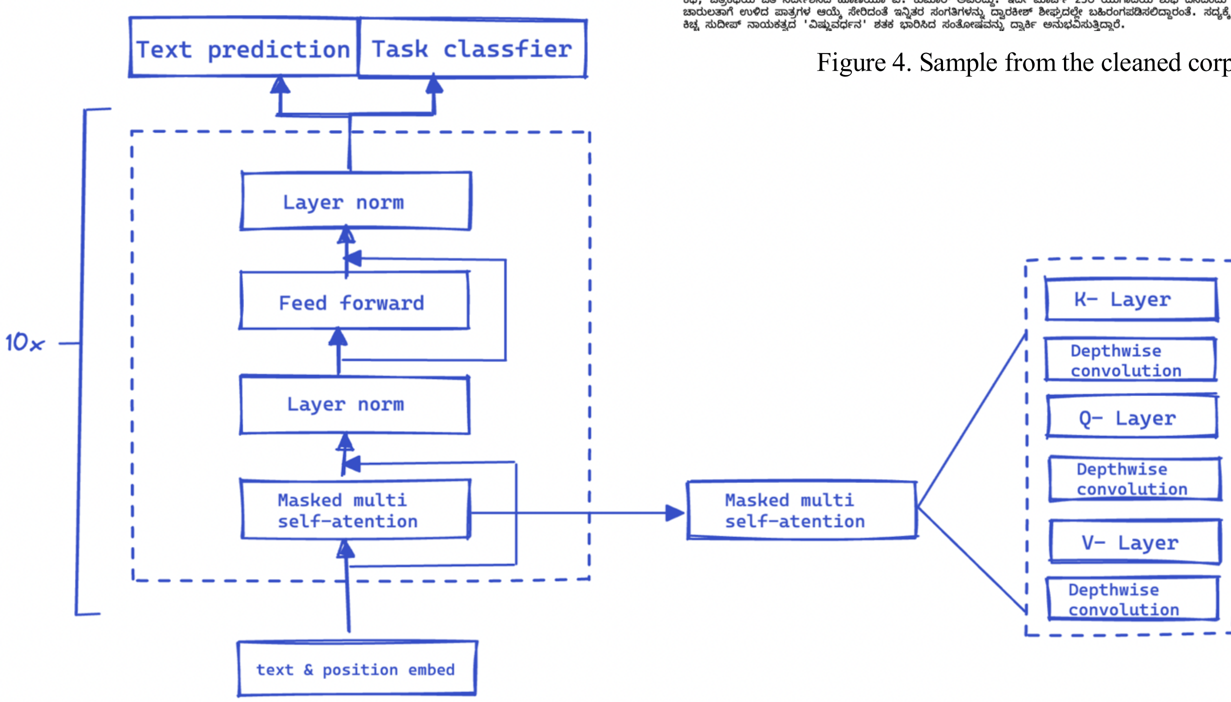

Figure 2. Improved GPT-2 architecture

C. *Pretraining and finetuning*

The pretrained GPT-2 model was used in our study. No additional pre-training was employed. The improved GPT-2 based model GPT-K was then trained on a large corpus of Kannada text. The training sequences were then fed into the transformer model to train an autoregressive model. Care was taken to prevent model overfitting [12] as it tends to retrieve raw sentences from the corpus.

D. *Training*

The model was trained with approximately 9.5 billion tokens for 10000 steps with a batch size of 1 for a single epoch. The corpus created in B was used as the dataset. The training takes approximately 150 hours on 12 intel i7 CPUs. A MiniForge3-based Conda environment was used to train the model implemented in TensorFlow. Both Adam[13] and SGD[14] optimizers were experimented with for compute optimization. SGD consistently showed better optimization results and resulted in lesser compute requirements than Adam. This is due to the lesser number of book-keeping variables in SGD than Adam [15]. A tensor rematerialization framework, as proposed in [16], is used for graph optimization to further reduce compute requirements. SGD shows better training and testing accuracies than Adam. Figure 5 shows the training and testing accuracies and losses for different optimizers where Adam is clearly outperformed by SGD. Top K sampling was used as the sampling method.

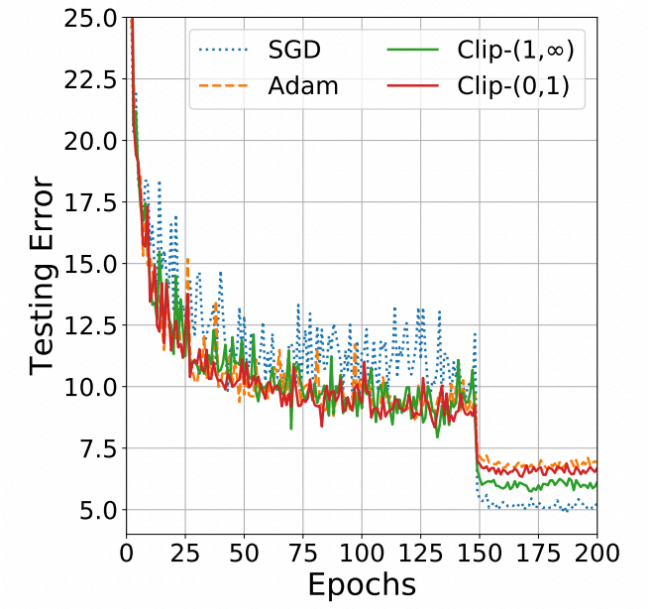

Figure 5. Comparison of Adam and SGD optimizers[15]

E. *Hyperparameter finetuning*

Finetuning hyperparameters showed a significant reduction in loss during model training. Increasing the learning rate showed a significant reduction in losses up to a point, followed by an exponential increase in the training loss. Increasing the number of attention heads and top k showed similar results, except that the loss linearly increased after a certain threshold.

Table 1. and Figure 6 show the correlation between learning rate and training loss. Table 2 and Figure 7 shows the correlation between top_k and learning loss. Table 3 and Figure 8 shows the relation between the number of attention heads and learning loss.

Learning rate	Loss
0.0001	1.1
0.0005	1.09
0.001	1.12
0.005	1.08
0.01	1.04
0.05	10.34
0.1	679

Table 1. Correlation of learning rate and loss

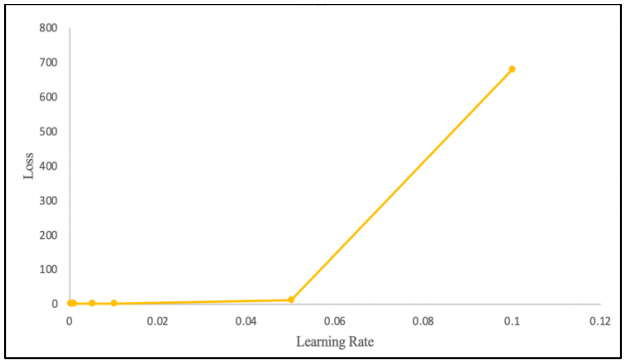

Figure 6. Correlation between learning rate and loss

Top_k	Loss
40	1.47
60	1.36
80	1.35
100	1.33
120	1.35
140	1.32
160	1.3
180	1.3
200	1.28
220	1.26
240	1.25
260	1.28

Table 2. Correlation between top_k and loss

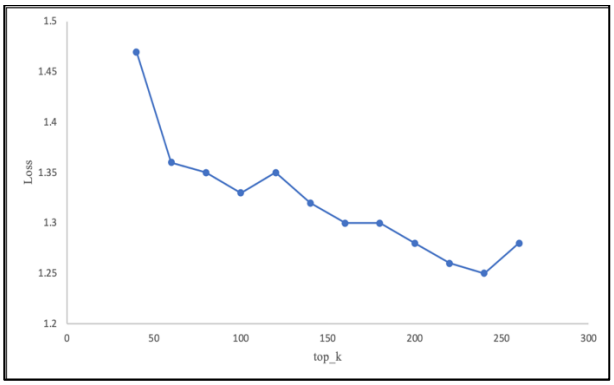

Figure 7. Correlation between top_k and loss

Number of attention heads	Loss
10	1.25
12	1.21
14	1.19
16	1.13
18	1.32

Table 3. Correlation between Number of attention heads and loss

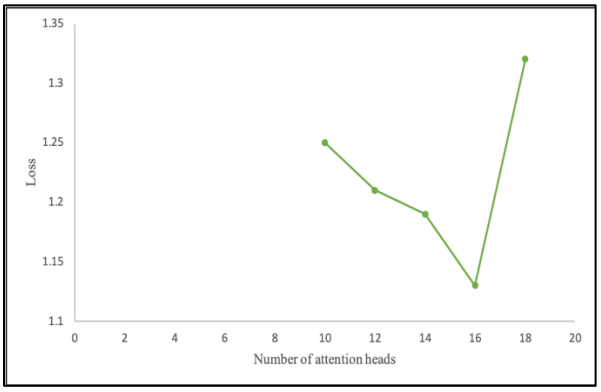

Figure 8. Correlation between Number of attention heads and loss

F. Evaluation

Finetuning Hyperparameters showed a significant reduction in loss during model training. Increasing the number of attention heads and layers significantly reduced the learning loss. The model was evaluated on one major standard benchmark called RecallOriented Understudy for Gisting Evaluation (ROUGE) [17]. Equation (2) is used to calculate the F-measure.

$$F - measure = 2 * \frac{precision * recall}{Precision + recall}$$

Equation 2.

IV. Generated Samples

The below images show certain samples of generated text which are both syntactically and semantically correct. We observed that the model generated syntactically correct text in most cases, but the generated text was not semantically correct except in rare cases. This can be attributed to short training times and low batch sizes. Figures 9 through 13 show the samples generated by the model.

```
Model prompt >>> ಕನ್ನಡ ರಾಜ್ಯೋತ್ಸವ
63 ನೇ ಕನ್ನಡ ರಾಜ್ಯೋತ್ಸವವನ್ನು ಸಡಗರ ಹಾಗೂ ಸಂಭ್ರಮದಿಂದ ಆಚರಿಸಲಾಯಿತು
```

Figure 9.

```
Model prompt >>> ಶಕ್ತಿಕಾಂತ ದಾಸ್
ಆರ್ ಬಿಐ ನೂತನ ಗವರ್ನರ್ ಶಕಿಕಾಂತ ದಾಸ್
```

Figure 10.

```
Model prompt >>> ಪ್ಲಾಸ್ಟಿಕ್ ಕವರ್
ರಾಜ್ಯ ಸರಕಾರ ವಿವಿಧ ರೀತಿಯ ಪ್ಲಾಸ್ಟಿಕ್ ಕವರ್, ಬ್ಯಾನರ್, ಬುಟ್ಟಿಗ್ಸ್ ಬಳಕೆ ನಿಷೇಧ ಆದೇಶ
```

Figure 11.

```
Model prompt >>> ಆಕ್ಟೋಪಸ್
ಆನೇಕ ಸಂಸ್ಕೃತಿಗಳಲ್ಲಿ ಮಾನವು ಆಕ್ಟೋಪಸ್ ಅನ್ನು ತಿನ್ನುತ್ತಾರೆ
```

Figure 12.

```
Model prompt >>> ಆಕ್ಟೋಪಸ್
ಆಕ್ಟೋಪಸ್ ಗಳು ತೆವಳುತ್ತಾ ಆಥವ ಈಜುತ್ತಾ ಚಲಿಸುತ್ತವೆ
```

Figure 13.

V. Future Work

The model can be further improvised to reduce compute requirements. Minor changes to the model architecture can be considered to this effect. Further, the model training time can be increased. The model can be trained on a larger corpus of data for more epochs with large batch sizes to improve its accuracy. Better optimizers can be used to improve compute efficiency. Further studies on hyperparameter finetuning can be considered to reduce loss and improve the efficiency of the model.

VI. Conclusion

Although the development of language models is taking place on a global scale, the use of AI assistants in regional languages has not yet been developed. Despite the fact that international markets are becoming more diverse, there is still an underdeveloped market for regional language models. The model presented in this paper is a GPT-2 based model, which can generate text in the Kannada language. Though the samples are not perfect all the time, the model is a good start for further research. We present this study in the hope that it will become a prototype for language models in regional languages.

REFERENCES

[1] Vaswani et al, 2017 “Attention is All you Need”, Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, 2017

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). “Improving language understanding by generative pre-training.”

[3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), p.9.

[4] Brown et al., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

[5] So, et al. "Searching for Efficient Transformers for Language Modeling." *Advances in Neural Information Processing Systems* 34 (2021): 6010-6022.

[6] Liao Y, Wang Y, Liu Q, Jiang X. Gpt-based generation for classical chinese poetry. arXiv preprint arXiv:1907.00151. 2019 Jun 29.

[7] C. R. Dhivyaa, K. Nithya, T. Janani, K. S. Kumar and N. Prashanth, "Transliteration based Generative Pre-trained Transformer 2 Model for Tamil Text Summarization," *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9740991.

[8] J Abadji, P O Suarez, L Romary, B Sagot, 2022, “Towards a Cleaner Document-Oriented Multilingual Crawled Corpus”, arXiv e-prints,

[9] Conneu et al., 2020 “Extracting High-Quality Monolingual Datasets from Web Crawl Data,” 2020, Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, 4003--4012

[10] Wikimedia Foundation, Wikimedia Downloads, wiki dump

[11] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. 2015 Aug 31.

[12] Hawkins, Douglas M. "The problem of overfitting." *Journal of chemical information and computer sciences* 44.1 (2004): 1-12.

[13] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[14] Ruder, S., 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

[15] Keskar, Nitish Shirish, and Richard Socher. "Improving generalization performance by switching from adam to sgd." *arXiv preprint arXiv:1712.07628* (2017).

[16] Kumar R, Purohit M, Svitkina Z, Vee E, Wang J. Efficient rematerialization for deep networks. Advances in Neural Information Processing Systems. 2019;32.

[17] Lin CY. Rouge: A package for automatic evaluation of summaries. InText summarization branches out 2004 Jul (pp. 74-81).