International Journal of Advance Research in Science and Engineering
Volume No. 14, Issue No. 04, April 2025
www.ijarse.com
IJARSE
ISSN 2319 - 8354

ContentBot: Generating Engaging Post for any Platform

Naman Sharma[1], Devraj Jajoo[2], Farhaan Ahmad[3], Ms. Adlene Ebenezer P[4]

[1,2,3] Student, [4]M.E., Assistant Professor, Dept. of Computer Science and Engineering,
SRM Institute of Science and Technology, Chennai, India

ABSTRACT

In today's fast-paced world where everyone consumes internet data specifically social content on daily basis, it has become crucial for the social media platforms to improve their platforms and in order to do that they can use ContentBot. ContentBot is an AI-power tool which will use the best models out there to generate interactive captions for your post so one can put out as many post one wants. The bot is going to use some of the most famous pretrained models and an LSTM for decoding the caption. The bot is also going to use the famous transformer models to generate captions but there will be no training for them. The evaluation for all the models will be done will the help of various ROGUE scores. In the output we notice is that KOSMOS and BLIP outperform all the other models by having the highest ROUGE score amongst all the models.

Keywords *— Image Captioning, CNN pre-trained models, Transformer-based models, Transformer pre-trained models, Transfer Learning*

1. INTRODUCTION

Content creation is one of the most common demands in the current world. We have seen number of people using and consuming social media content have increased significantly over the past 5 years. Meta, one of the most renowned companies for the creation of Facebook and Instagram has about 95 million photos and videos shared on daily basics. Also, the user base of Instagram in India is over 362.9 million users which is the highest in the world. They have over 1,127,000 Instagram influencers. These Influencers post at least one photo per day which is not easy task to do on a daily basics. In order to help making the life of an influencer much easier creation of the ContentBot was done. Also, the same company has developed AI based virtual accounts (Ai controlled accounts) which can also use the bot for caption generation. The bot is trained in such a way that it will extract some important features from the image and based on that it will generate a caption. There will be two approached used first one being Transfer Learning and second one being the Transformer usage.

Training a fresh Deep Learning model could require machine with specific high-end computation, which is not readily available for everyone so, Transfer Learning was developed as the solution. Transfer Learning (TF) is a technique used in machine learning which allows a coder to utilize a pre-trained model (model which has been already trained on some data and gives decent results) to develop a solution for the problem at hand. Anderw Ng, one of the creators of the TensorFlow library said that after supervised learning transfer learning will the way ahead for generating solution in the field of AI. The reason to use Transfer Learning is that TF reduces the time which will be required for training the model and also having a good pre trained model can enhance the outcome. Content Bot utilizes image based pre-trained models like Resnet-50, VGG16, DenseNet, and InceptionV3.

Back in 2017, some researchers at google generated a new model called the Transformer model which was used

for the sequence to sequence based natural language processing task (NLP). This research revolutionized the AI industry as most of the AI model we see today like GPT, Gemini, Claude etc. are made using the same architecture. This inspired researchers to create such model which use transformer model for computer vision-based problem. Over time many different models are created and the bot uses some of them like vit-gpt2(combined model), BLIP model by salesforce. We will test the models and choose the most suitable one.

2. LITERATURE SURVEY

Automated image captioning has seen a very visible spike in interest. This is likely because of the crucial role social media plays in the influencer market. A good caption makes a world of difference and can help influencers stand out from the rest of the competition. From its inception, social media has seen a steady rise in influencers and the various different content they make. As social media reaches a more global audience, the need for good captions is greater than ever to improve engagement, accessibility and break cultural and linguistic barriers to convey messages to a much more diverse audience. Research has been going on to find out optimal ways in which the caption generation process can be automated so that the influencers can focus their energies on more creative pursuits.

Early research focused on utilizing AI for automated caption writing to streamline content creation. Roshne et al. [1] took a look at the role of Recurring Neural Networks (RNNs) and the Natural Language Processing (NLPs) in generating relevant, contextually sensible captions. Their findings lead to huge breakthroughs in the field of caption generation. Their findings led to huge breakthroughs in the field of caption generation. Later advancements integrated transformer-based architectures such as BERT and GPT, significantly improving caption accuracy and coherence. Additionally, multimodal approaches, combining image processing models like CNNs and DenseNet with NLP techniques, further refined the contextual understanding required for high-quality automated captioning.

Building on this, Yahya et al. [2] made use of Convolutional Neural Networks (CNNs) to generate captions and hashtags. Furthermore, they integrated attention mechanisms to refine caption quality and improve coherence. Similarly, Raju et al. [3] made use of a framework for generating images and hashtags for social media posts. The AI framework makes use of diffusion models and interactive evolution computation to generate engaging social media content. Tested on Instagram, it adapts to user feedback, delivers high-quality content, and sustains audience engagement, showcasing the potential of artificial influencers in digital marketing.

More advanced techniques have been employed to enhance caption accuracy. Zhang et al. [4] integrated CNNs for feature extraction with Long Short-Term Memory (LSTM) networks for sentence generation, significantly improving the precision of generated captions. Combining computer vision and NLP, it creates accurate image captions. Tests show its effectiveness in generating precise and natural descriptions. Syed et al. [5] contributed by further advancing this by combining VGG-19 with LSTM, making good utility of feature extraction and sequence modeling to produce high-quality captions. Their model, trained on the Flickr8k dataset, achieved a BLEU score of 0.669135, highlighting its effectiveness.

Many other studies have utilized the encoder-decoder architectures with attention mechanisms to optimize captioning models. U.Kulkarni et al. [6] made use of an attention-based approach with CNN and Gated Recurrent Unit (GRU) networks, improving caption accuracy by 10% over traditional methods. The study incorporated

Bahdanau's attention mechanism. It made use of InceptionV3 for image feature extraction and GRU for caption generation. Being trained in the Flickr8k dataset, it generates more accurate and detailed captions by giving attention to important parts of the images.

Yang et al. [7] introduced an LSTM-GAN framework, improving on error accumulation issues in traditional LSTM-based models and showing a superior performance in caption accuracy. Experimental results show that this LSTM-GAN architecture outperforms existing video captioning methods on standard datasets. More recent studies continue to explore the synergy between CNNs, RNNs, and attention mechanisms. Zheng-Jun Zha et al. [8] came up with an interesting concept called Context-Aware Visual Policy Network (CAVP) which greatly enhanced the image captioning process by taking into account both the past and future visual information. Unlike the usual methods in which caption generation takes place in a sequential manner without much context, CAVP uses a structured approach that adjusts the caption dynamically. This provides a great deal of versatility. This improves semantic comprehension by integrating context within images, making it more accurate as a result

Oriol Vinyals et al. [9] came up with the show and tell model. It's a deep learning model approach to image caption generation. The results were groundbreaking, leveraging the amazing power of deep learning to come up with effective and and coherent text generation. Because of the powerful neural networks used in this model, the accuracy of the generated captions has increased tremendously. CNN and RNNs for multi-modal processing, effectively bridging the gap between visual and textual data to create cohesive and context-aware captions. Finally, Longwu Yang et al. [10] developed CaptionNet, a recurrent neural network (RNN) designed to improve image caption generation. CaptionNet improves on pre-existing LSTM inspired captioning by adding attention mechanisms, which allows for selectively focusing on the features of the image which matter the most. By taking this crucial addition to the Caption Net architecture, it outperforms base models and results in improved accuracy for image caption generation. Applications of this model include content creation for various social media platforms, and other practical applications which involve caption generation

3. DESIGN CONSIDERATION

To create image captioning model, understanding the architecture, the approach and data is an essential step.

3.1. Dataset for Transfer Learning models:

The famous Flickr 8k dataset has been selected for TF model. The dataset was downloaded from Kaggle, a website where many datasets are available for free of cost. The dataset consists 8091 images and for each image there are 5 human generated captions stored in a comma separated file (.csv file). The image are colored images of 3 channels (RGB). The following images (Fig. 3.1 to Fig. 3.3) shows some of the images and one of the corresponding captions. The training dataset is 85% of the dataset and the rest 15% is the validation dataset.

Fig. 3.1: Sample Image 1

Fig. 3.2: Sample Image 2

Fig. 3.3: Sample Image 3

3.2. Architecture Diagram

The following image (Fig. 3.4) shows the architecture diagram for the TF based model. The models we are going to use are Resnet-50, VGG16, DenseNet, and InceptionV3. We are applying early stopping while training these models.

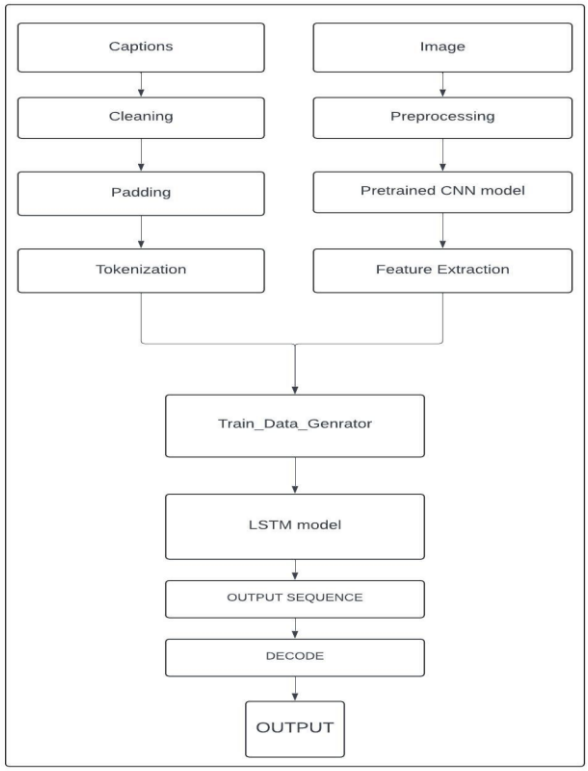

Fig. 3.4: Architecture Diagram

3.3. Cleaning / Padding and Pre-processing:

In the above figure (Figure 1.3) the cleaning process done on the caption consist of the following process. Lowering – the text was made into lower case. Removing extra characters – the numbers and special characters from the text are removed using regex. Removing spaces – the intermediate space is removed. Conversion – the previous step was done by making the text into list. Now, it is converted back to string. As for the preprocessing of image, there is only one thing we need to do that is convert into right size.

3.4. Dataset of Transformer based Models:

The transformer-based models are pre-trained on the datasets like MS COCO famously known as the COCO dataset etc. So, 1000 images will be tested on these models and their BLEU score will be taken in consideration. Based on the outcome the top 2 models will be used for the bot.

3.5. Architecture for Transformer based models:

The figure below (Fig. 3.5) represents the architecture diagram for the caption generating process using the transformer-based models.

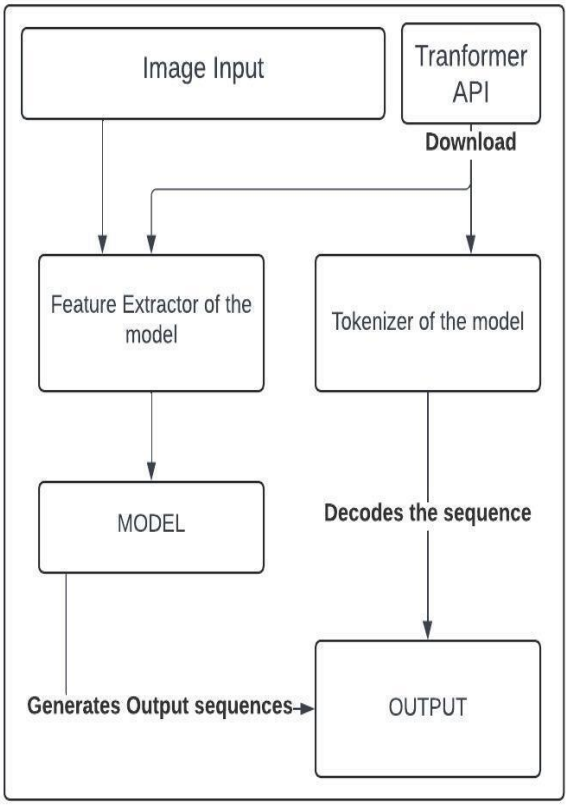

Fig. 3.5: Architecture Diagram

We need to install the tokenizer and feature extractor before we move on to use the models, the transformer library from Hugging Face is used as all the models used for this research purpose are taken from the Hugging face website. The models used are mentioned below: -

3.5.1 BLIP model

An image caption model trained on ImageNet (from ViT), COCO and Visual Genome datasets. This was made

the Salesforce and is one of the most famous models used for this purpose. In this research we are using its base version of this model. (1)

3.5.2 KOSMOS model

It is a fine-tuned model made by Microsoft. It was originally a model which was used for text-to-text pro (2) but the version used in the research has been fine-tuned for the image captioning purpose.

3.5.3 GIT model (3)

It is another model made by Microsoft which is a decoder only model having 177 million parameters. We are using the base version of this model.

3.5.4 ViT-GPT2 model

It is a model which is used by combining ViT and GPT2 model. The GPT2 model doesn't poses any image processing techniques so it is combined with ViT for feature extraction.

4. PERFORMANCE ANALYSIS

4.1 Setting up the basic Information

For all the models we are going to check ROGUE scores and test the model on 50 images of No-Cap Dataset. The reason to use No-Cap is that both Transformer-based and TF models are not trained on this data. The following image (Fig. 4.1) will be used for human level comparison of the captions generated by the models. The averages of the model as calculated and graphed as well.

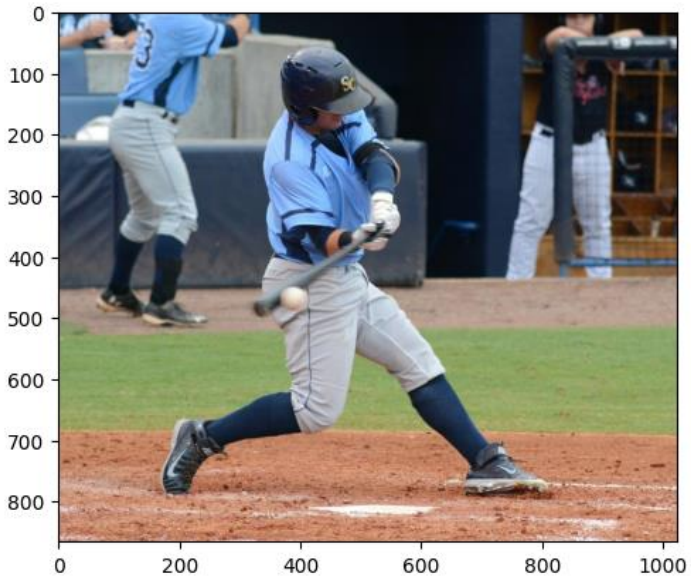

Fig. 4.1: Test Image

4.2 ROUGE Scores

The Rouge scores used in this research are as follows: -

4.3.1 ROUGE – 1 SCORE:

The rogue - 1 score is the most basic of all the rouges scores out there. In-order to calculate the rogue-1 score we need to calculate the precision and recall for the same. The formula for calculating the rouge - 1 score is as follows:

$$R1\ Precision\ (P) = \frac{unigrams\ matching\ in\ reference\ and\ output}{total\ no.\ of\ unigrams\ in\ the\ output}$$

$$R1\ Recall\ (R) = \frac{unigrams\ matching\ in\ reference\ and\ output}{total\ no.\ of\ unigrams\ in\ the\ reference}$$

$$ROUGE\ 1\ Score\ (R1) = 2 * \frac{P * R}{P + R}$$

Here the reference is the human given caption and the output is the model generated output.

4.3.2 ROUGE – 2 SCORE:

The rouge – 2 score is similar to that of rouge – 1 score, the only difference is that we are comparing two words (bigrams) at a time instead of one word (unigrams) as done in rouge – 1. The formula to calculate the rouge - 2 score is as followed: -

$$R2\ Precision\ (P) = \frac{bigrams\ matching\ in\ reference\ and\ output}{total\ no.\ of\ bigrams\ in\ the\ output}$$

$$R2\ Recall\ (R) = \frac{bigrams\ matching\ in\ reference\ and\ output}{total\ no.\ of\ bigrams\ in\ the\ reference}$$

$$ROUGE\ 2\ Score\ (R2) = 2 * \frac{P * R}{P + R} \quad (4)$$

4.3.3 ROUGE - L SCORE: (5)

Many times, the rouge-1 or rouge-2 score do not yield the best results so to solve that we use rouge – L where L stands for Longest Common Subsequence (LCS). It checks for the length of the longest common words (6)

$$RL\ Precision\ (P) = \frac{LCS(generated\ , reference)}{total\ no.\ of\ bigrams\ in\ the\ output}$$

$$RL\ Recall\ (R) = \frac{LCS(generated\ , reference)}{total\ no.\ of\ bigrams\ in\ the\ reference}$$

$$ROUGE\ L\ Score\ (RL) = 2 * \frac{P * R}{P + R} \quad (7)$$

4.3 Transfer Learning Models

Now we will look how the Transfer Learning (TF) models – Dense Net, Resnet50, VGG16 and Inception V3 got trained and performed on Flick 8K dataset. (8)

4.3.1 Dense Net

The Dense Net model got trained for 18 epochs before the early stopping took place. The training loss is: 3.0323 and the validation loss is: 3.55084. The following figure (Fig. 4.2) shows the training and validation loss per epoch. (9)

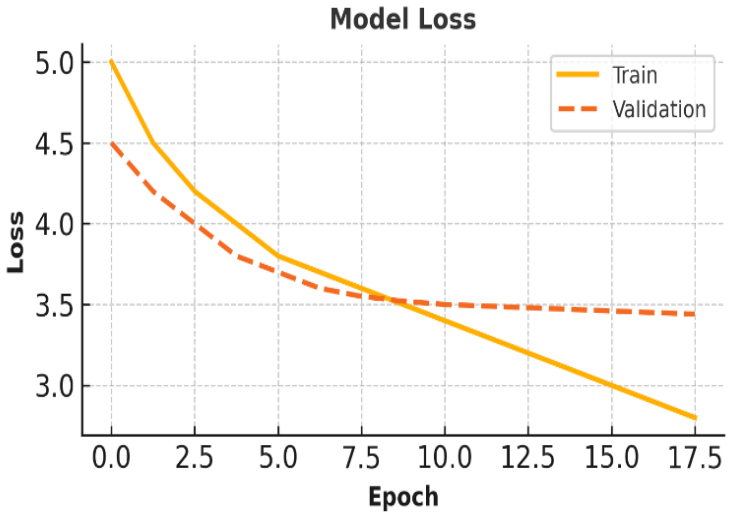

Fig. 4.2: Dense Net Loss Graph

Now let's see what output did Dense Net generated for the Test Image (Fig. 4.1).

Output: 'boy in blue shirt is swimming in the water'

From the output we can tell that it is not at all correct so, we can say the model as gotten overfit as it cannot answer correctly to the test image.

The best rouge score achieved were against Reference- caption number 5 'A person in blue is swinging a baseball bat at the ball'. Rouge-1, Rouge-2 and Rouge-L score for the same are 0.3478, 0.09523, 0.3478. Rouge-1 & Rouge-L are equal so the meaning of this is that the LCS is of 1 word or unigrams. The overall performance of the Dense Net is given in the following image (Fig. 4.3): -

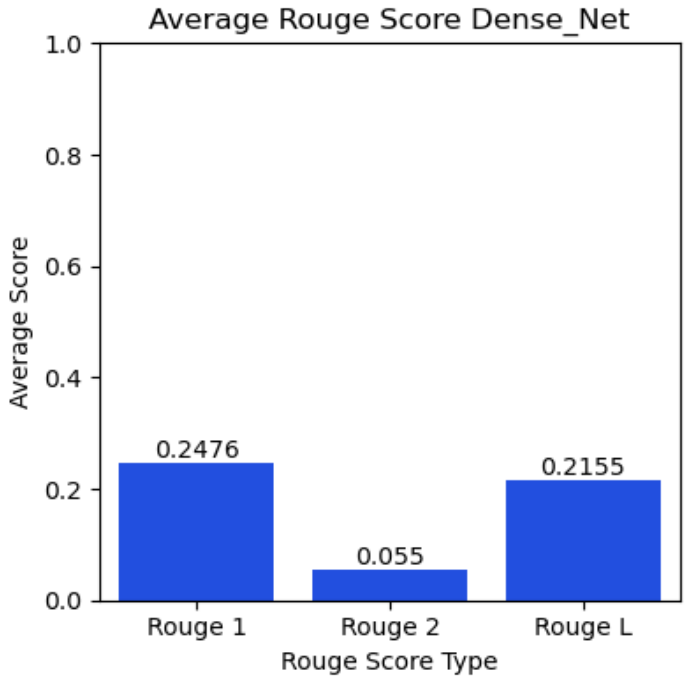

Fig. 4.3: Dense Net Average Rouge Graph

4.3.2 Resnet 50

The Resnet50 model got trained for 27 epochs before the early stopping took place. The training loss is: 3.3416 and the validation loss is: 3.425. The following figure (Fig. 4.4) shows the training and validation loss per epoch.

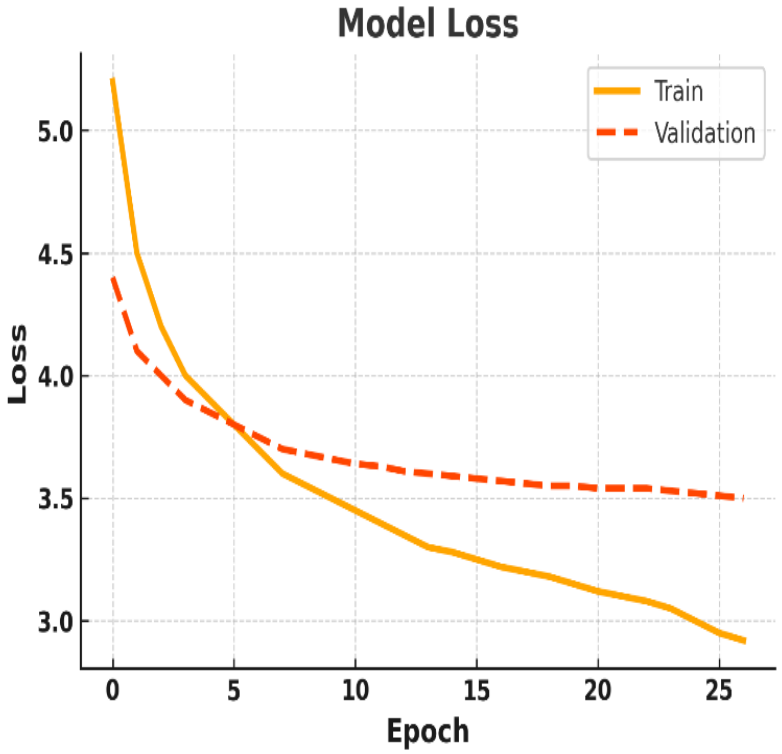

Fig. 4.4: Resnet 50 Loss Graph

Now let's see what output did Resnet generated for the Test Image (Fig. 4.1).

Output: 'two dogs are running through the grass'

From the output we can tell that the model is completely off so, we can say the model as gotten overfit as it cannot caption any image.

The best rouge score achieved were against Reference- caption number 3 'The boy is playing baseball along the baseball field'. Rouge-1, Rouge-2 and Rouge-L score for the same are 0.1111,0.0,0.1111. Rouge-1 & Rouge-L are equal so the meaning of this is that the LCS is of 1 word or unigrams.

The overall performance of the Renset50 is given in the following image (Fig. 4.5): -

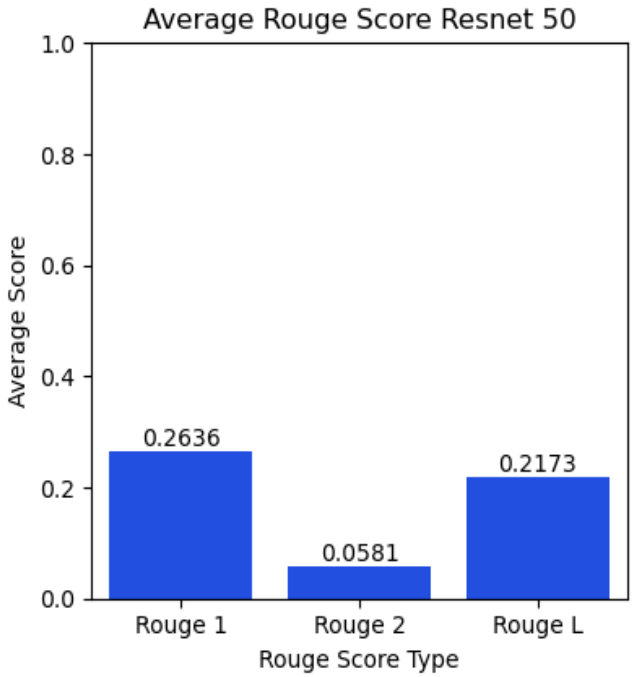

Fig. 4.5: Resnet 50 Average Rouge Graph

4.3.3 VGG16

The VGG16 model got trained for 24 epochs before the early stopping took place. The training loss is: 3.1607 and the validation loss is: 3.6463. The following figure (Fig. 4.6) shows the training and validation loss per epoch.

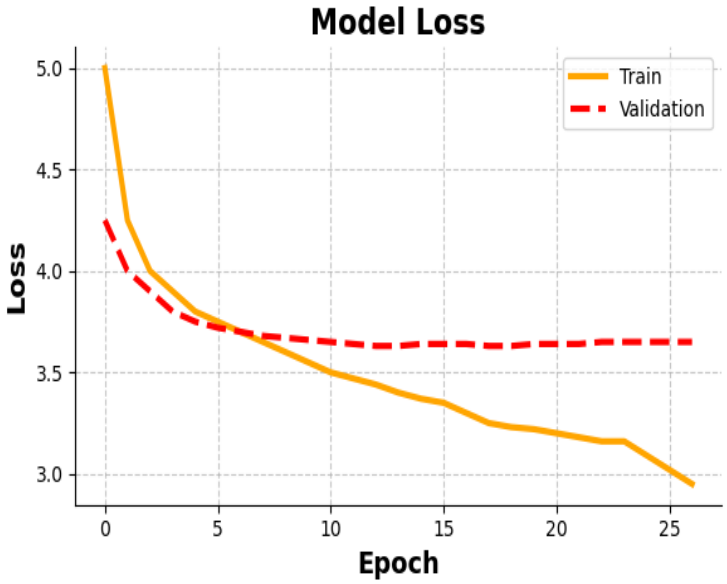

Fig. 4.6: VGG16 Loss Graph

Now let's see what output did VGG16 generated for the Test Image (Fig. 4.1).

Output: 'man in blue shirt is climbing rock'

From the output we can tell that it is not at all correct so, we can say the model as gotten overfit as it cannot answer correctly to the test image.

The best rouge score achieved were against Reference- caption number 5 'A person in blue is swinging a

baseball bat at the ball'. Rouge-1, Rouge-2 and Rouge-L score for the same are 0.2857,0.1052,0.2857. Rouge-1 & Rouge-L are equal so the meaning of this is that the LCS is of 1 word or unigrams.

The overall performance of the VGG16 is given in the following image (Fig. 4.7): -

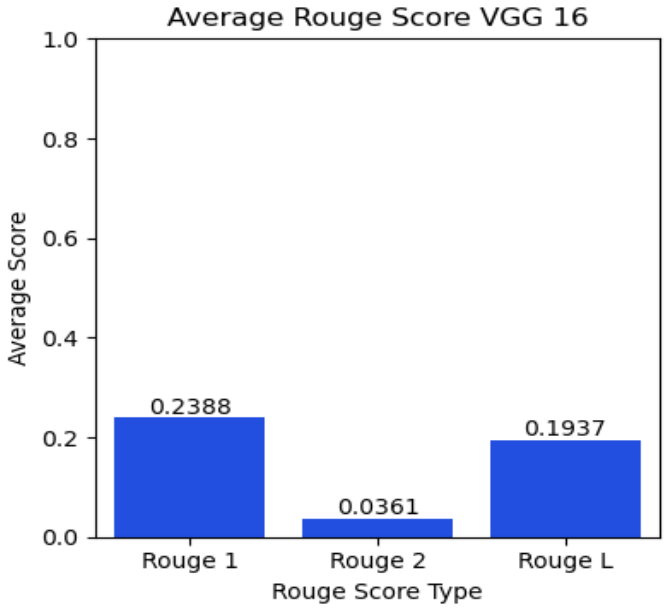

Fig. 4.7: VGG16 Average Rouge Graph

4.3.4 Inception V3

The Inception V3 model got trained for 17 epochs before the early stopping took place. The training loss is: 2.5235 and the validation loss is: 2.9719. The following figure (Fig. 4.8) shows the training and validation loss per epoch.

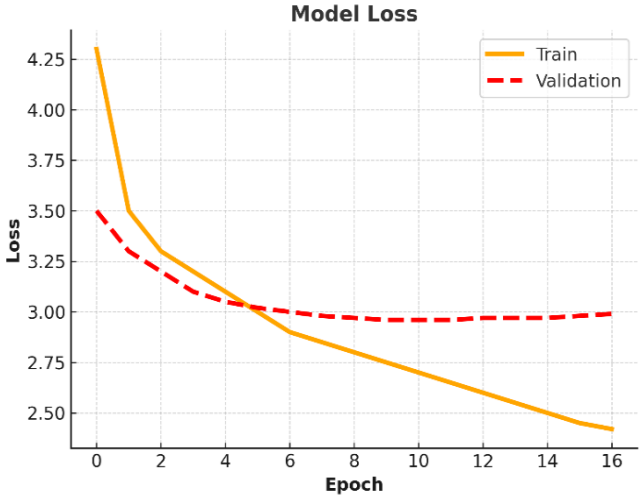

Fig. 4.8: Inception V3 Loss Graph

Now let's see what output did VGG16 generated for the Test Image (Fig. 4.1).

Output: 'boy in blue shirt is playing in the water'

From the output we can tell that it is not at all correct so, we can say the model as gotten overfit as it cannot answer correctly to the test image.

The best rouge score achieved were against Reference- caption number 3 'The boy is playing baseball along the

baseball field'. Rouge-1, Rouge-2 and Rouge-L score for the same are 0.3999,0.1111,0.3999. Rouge-1 & Rouge-L are equal so the meaning of this is that the LCS is of 1 word or unigrams.

The overall performance of the Inception V3 is given in the following image (Fig. 4.9): -

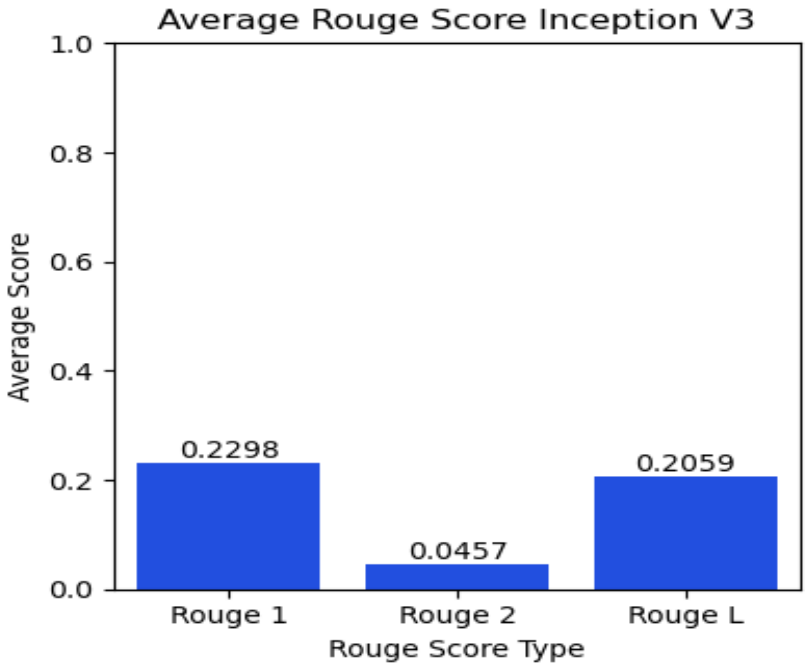

Fig. 4.9: Inception V3 Average Rouge Graph

4.4 Transformer Models

Now let's see how does Transformer-based perform on the Test image (Fig. 4.1). The models are not trained as they are already pre-trained. The models used are – ViT-GPT2, BLIP, GIT and Kosmos.

4.4.1 ViT-GPT2

The ViT-GPT2 model is a model made with combination of two models. The first one is Vision Transformer (ViT) made by Google which is trained on ImageNet dataset and GPT2 by OpenAI which is a decoder-only model which it was used in the famously known AI ChatGPT. Since GPT2 wasn't made to handle image processing tasks so, ViT was added to for image feature extraction purpose. Now let's see the output of the model on the Test image (Fig. 4.1).

Output: 'a baseball player swinging a bat at a ball'

As we can see that the output is very relatable to the image and can be considered a viable output for a problem. The Rouge-1, Rouge-2, and Rouge-L score for this output are 0.7999, 0.6666, 0.7999 respectively. The scores are in reference to the caption number 2 – 'A baseball player swinging his bat at a ball'. The overall performance of the model is in the figure (Fig. 4.10): -

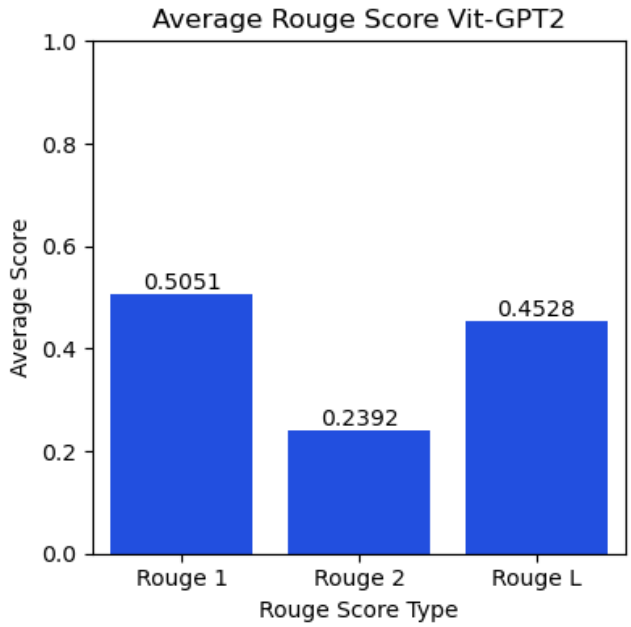

Fig. 4.10 ViT-GPT2 Average Rouge Graph

4.4.2 BLIP – Bootstrapping Language-Image Pre-Training

The BLIP model is most famous for image captioning. This model has been trained over 14 M images on the base version. The large version of this model is trained on 129 M images. These images come from various dataset like COCO, Visual Genome, Conceptual Caption, SBU captions, LAION. This research uses the base model as due to the restriction of compute requirements. Now let's see how the model performs on the Test Image (Fig. 4.1).

Output: 'a baseball player swinging a bat on a field'

The output generated the model is very accurate and has the following Rouge-1, Rouge-2 and Rouge-L score: 0.6,0.3333,0.6. These scores are in represent to the caption number 2- 'A baseball player swinging his bat at a ball'. The overall performance of the model is given in the image below (Fig. 4.11): -

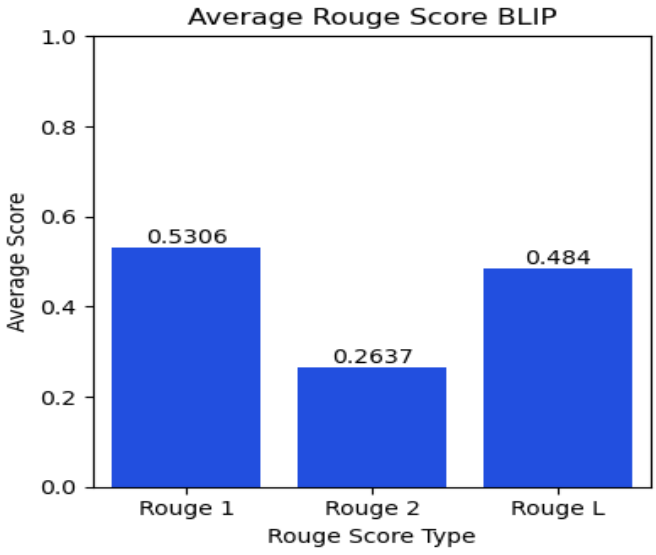

Fig. 4.11 BLIP Average Rouge Graph

4.4.3 GIT - GenerativeImage2Text

It is a was developed as decoder only model and it utilizes CLIP model image tokens and text tokens. This model was trained on the following image-caption datasets: COCO, SBU, Visual Genome, Conceptual Captions and ALT200M. It was trained on over 800M images making it the biggest model for image captioning purpose. This model can also do video-captioning tasks as well. Now let's see how the model performs on the Test Image (Fig. 4.1): -

Output: 'a baseball player swinging a bat'

The output generated the model is very accurate and has the following Rouge-1, Rouge-2 and Rouge-L score: 0.8,0.4615,0.6667. These scores are in represent to the caption number 2- 'A baseball player swinging his bat at a ball'. The overall performance of the model is given in the image below (Fig. 4.12): -

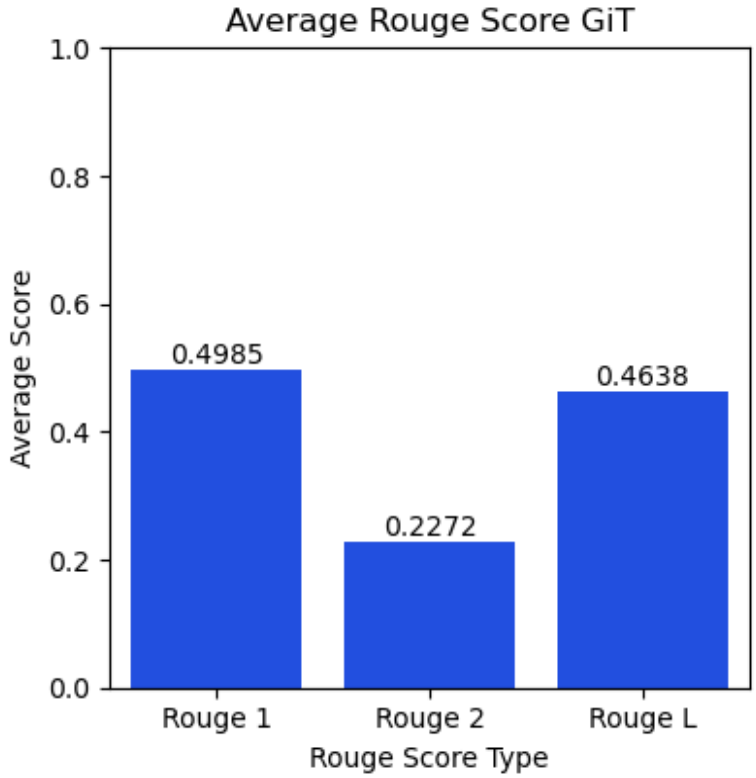

Fig. 4.12 GIT Average Rouge Graph

4.4.4 KOSMOS

The original KOSMOS-2 model is an LLM known for its image related tasks handling and the version used for this research is the patch-144 version of the models which is an image-to-text model and is also developed by Microsoft. The output always has 'An image of' in the start Not much information is available for its training on the image captioning task so, now let's see how it performs on the Test Image (Fig. 4.1): -

Output: 'An image of a baseball player swinging a bat and ball'

The output generated the model is very accurate and has the following Rouge-1, Rouge-2 and Rouge-L score: 0.7368, 0.3529, 0.6315. These scores are in represent to the caption number 2- 'A baseball player swinging his bat at a ball'. The overall performance of the model is given in the image below (Fig. 4.13): -

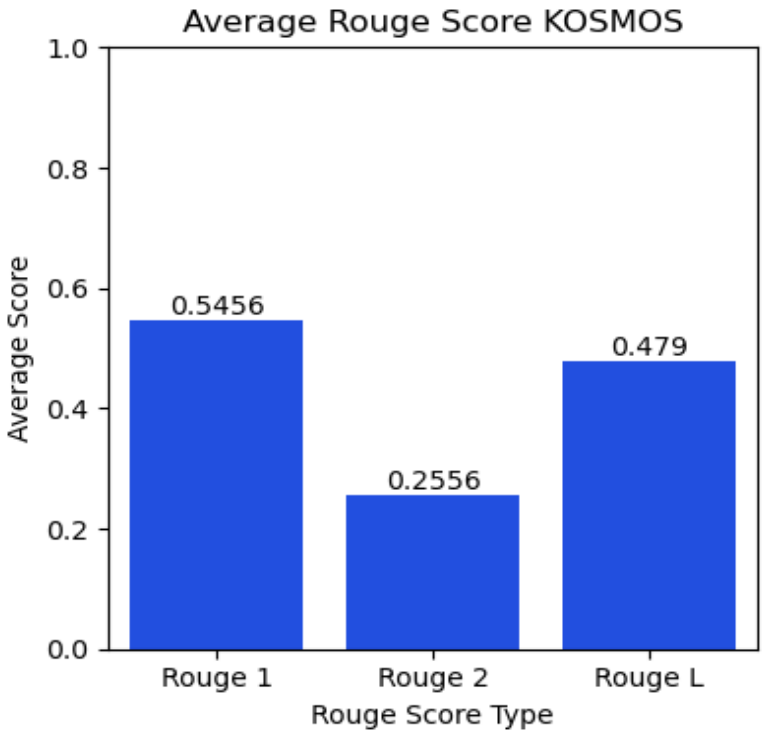

Fig. 4.13 KOSMOS Average Rouge Graph

5. Conclusion & Future Scope

After looking at the all the outcomes from all the models. It can be said that the transformers-based model performs exceptionally well in comparison to the Transfer Learning based models. This performs gap is due to the image dataset size used to train these models, Transformer based models are trained on dataset having millions of images and captions whereas the Transfer Learning models are only trained on few thousand images. In-order to decide which 2 models will be a part of ContentBot, the following graph will help in that (Figure 5.1):-

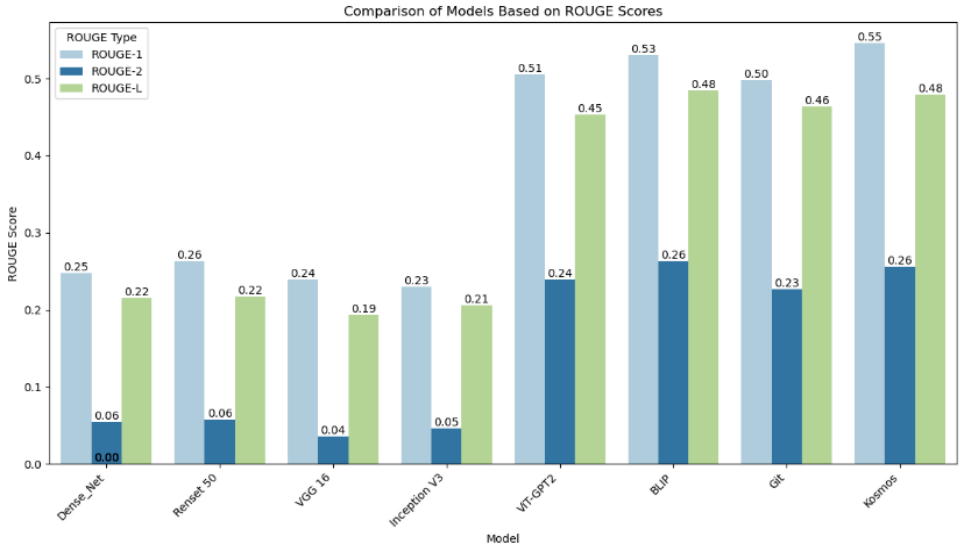

Fig5.1 Combined Results

It is clear that the model KOSMOS and BLIP out perform all the other models and hence would be the best models to be used in out ContentBot.

The future scope for this research is that the Transfer Learning model could be trained on dataset with more images as this is computationally expensive and also a constraint in our research. As new and better LLMs keep rolling out replacing the current used models with the better models will be another future scope of this research. One could try to fine-tune a LLM on a better dataset which will lead to a better LLMs for the same cause will be another future scope of this research.

REFERENCES

[1] R. V, B. S. R, S. V. R S and V. C, "Empowering Content Creation using Artificial Intelligence - The Role of Caption Writing: An Overview," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1-6.

[2] Y. Q. AL-Sammarraie, K. AL-Qawasmi, M. R. AL-Mousa and S. F. Desouky, "Image Captions and Hashtags Generation Using Deep Learning Approach," 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), Zarqa, Jordan, 2022, pp. 1-5,

[3] R. Shrestha and H. Korneliussen, "A Framework for Generating Images and Hashtags for Social Media Posts for Artificial Influencers," 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2024, pp. 42-48,

[4] C. Zhang, Y. Dai, Y. Cheng, Z. Jia and K. Hirota, "Recurrent Attention LSTM Model for Image Chinese Caption Generation," 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Toyama, Japan, 2018, pp. 808-813.

[5] S. Abudhagir U, K. Vignesh, U. Harish, V. Avinash and M. V. Subbarao, "CaptionCraft: VGG with LSTM for Image Insights," 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST), Kochi, India, 2024, pp. 1-6.

[6] U. Kulkarni, K. Tomar, M. Kalmat, R. Bandi, P. Jadhav and S. Meena, "Attention based Image Caption Generation (ABICG) using Encoder-Decoder Architecture," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1564-1572,

[7] Y. Yang et al., "Video Captioning by Adversarial LSTM," in IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5600-5611, Nov. 2018,

[8] R. V, B. S. R, S. V. R S and V. C, "Empowering Content Creation using Artificial Intelligence - The Role of Caption Writing: An Overview," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1-6.

[9] Y. Q. AL-Sammarraie, K. AL-Qawasmi, M. R. AL-Mousa and S. F. Desouky, "Image Captions and Hashtags Generation Using Deep Learning Approach," 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), Zarqa, Jordan, 2022, pp. 1-5

[10] Y. Yang et al., "Video Captioning by Adversarial LSTM," in IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5600-5611, Nov. 2018