Image Caption Generator

Ashwin Sadanandan Nambiar, Hemant Gautham Bhodare Guide: Prof Divya, Premchandran
Keraleeya Smajam's Model College, khambalpada Road Thakurli, Dombivili (East)

Abstract—— As there has been a rise in the use of AI for various tasks to simplify human work and to increase the accuracy of the tasks being performed, certain improvisation must be implemented. This project aims to implement an image caption generator that will generate a caption for the provided image. The ultimate purpose of this project is to enhance the user experience by generating automated captions. In this project, based on the image given will generate a caption from our model. The idea is that we will get an automated caption when we implement it on social media or any other platform. And can also be implemented in various other sectors of everyday life

Keywords Caption Generator, Machine Learning, Automated Captions, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)

I.INTRODUCTION

Generating accurate captions for an image has remained as one of the major challenges in Artificial Intelligence with plenty of applications ranging from robotic vision to help the visually impaired. Long-term applications also involve providing accurate video captions in scenarios such as security systems. "Image caption generator": the name itself suggests that we aim to build an optimal system that can generate semantically and grammatically accurate captions for an image. Re- searchers have been involved in finding an efficient way to make better predictions, therefore we have discussed a few methods to achieve good results. We have used deep neural networks and machine learning techniques to build a good model. We have used Flickr 8k dataset which contains around 8000 sample images with their five captions for each image. There are two phases: feature extraction from the image using Convolutional Neural Networks (CNN) and generating sentences in natural language based on the image using Transformer. For the first phase, rather than just detecting the objects present in the image, we have used a different approach to extracting features of an image which will give us details of even the slightest difference between two similar images. For the second phase, we need to train our features with captions provided in the dataset and t Transformer shave been used int his phase of operation

2.LITERATURE REVIEW

Many existing studies have been conducted in the field which will help us in enhancing the project. This section elaborates on the recent studies and research on the technology. They emphasize the role of captioning in various fields and their application. The research has been conducted in such a way as to design and generate a system that will match our objectives in the project.

[1] Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge 2015
This research presented a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. The model these researchers were able to produce was accurate and was verified both qualitatively and quantitatively.

Task-Adaptive Attention for Image Captioning
This paper point out how most attention models only focus on visual features. When generating syntax-related words, little visual information is needed. In this case, these attention models could mislead the word generation. In this paper, we propose a Task-Adaptive Attention module for image captioning, which can alleviate this

misleading problem and learn implicit non-visual clues which can be helpful for the generation of non-visual words

[2]A Systematic Literature Review on Image Captioning

This study a comprehensive Systematic Literature Review (SLR) provides a brief overview of improvements in image captioning over the last four years. The main focus of the paper was to explain the most common techniques and the biggest challenges in image captioning and to sum- marize the results from the newest papers. Inconsistent comparison of results achieved in image captioning was noticed during their study and hence the awareness of incomplete data collection is raised in this paper. Therefore, the researcher put forth the importance of comparing the results of a newly created model produced with the newest information and not only with the state-of-the-art methods. This SLR provides a source of such information for researchers in order for them to be precisely correct on result comparison before publishing new achievements in the image caption generation field.

[3]Adversarial Semantic Alignment for improved image captioning

The research focuses on the study of image captioning as a conditional GAN training, proposing both a context-aware LSTM captioner and a co-attentive discriminator, which enforces semantic alignment between images and captions. It shows that surprisingly, SCST (self-critical Sequence Training) (a policy gradient method) shows more stable gradient behavior and improved results over Gumbel ST, even without accessing the discriminator gradients directly. The research also addresses the open question of automatic evaluation for these models and introduces a new seman- tic score and demonstrates its strong correlation to human judgment. This research shows that the SCSST when implemented on the MSCOCO Dataset was able to produce a strong performance in both semantic score and human evaluation.

[4]Boosting Image Captioning with Attributes

In this research Long Short-Term Memory with Attributes (LSTM-A) - a novel architecture that integrates attributes into the successful Convolutional Neural Networks (CNNs) plus Recurrent Neural Networks (RNNs) image captioning framework, by training them in an end-to-end manner. Particularly, the learning of attributes is strengthened by integrating inter-attribute correlations into Multiple Instance Learning (MIL) is presented. To incorporate attributes into captioning, the researchers construct variants of architectures by feeding image representations and attributes intoRNNs in different ways to explore the mutual but also fuzzy relationship between them

[5]Conclusion

From past research, it can be concluded that image captioning is still in the phase of development as new and innovative ways of captioning are being explored. The most used captioning method that has been able to produce stable results in real-life scenarios would be CNN-RNN-based methods. After analyzing the research, it can be sure that with the increase in computing power and the development of Al the application of image caption technology has been made in security, social media platform, and many more. But the major problem is that most of the models producedfrom the research can give semantically right but grammatically wrong captions which are wrong according to human evaluation. For this research, research papers [4] and [5] will be consideredas our base papers as they provided the required information regarding the model to be generated and the methods most suitable for it.

3.METHODOLOGY

The model architecture used a 2-layer Transformer-decoder. To get the most out of this model we had to tune with text generation, seq2seq models attention, or transformers. The model architecture built in this tutorial is shown below. Features are extracted from the image and passed to the cross-attention layers of the Transformer-decoder.

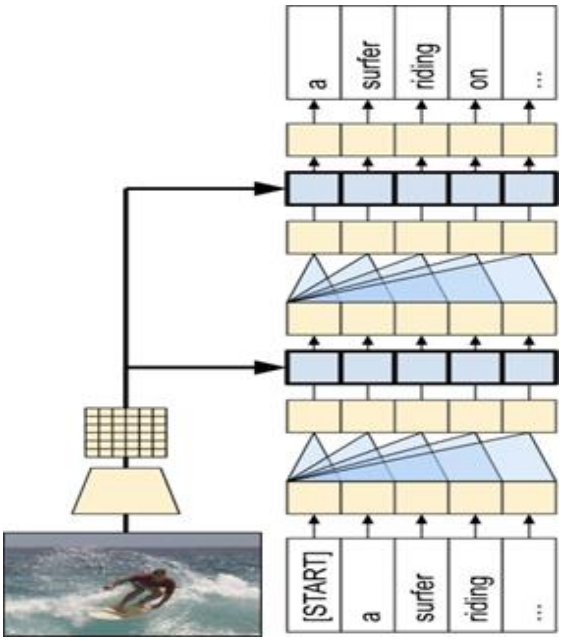

Figure 1: model architecture

3.1A Transformer encoder decoder model

This model assumes that the pre-trained image encoder is sufficient, and just focuses on building the text decoder. This tutorial uses a 2-layer Transformer-decoder.

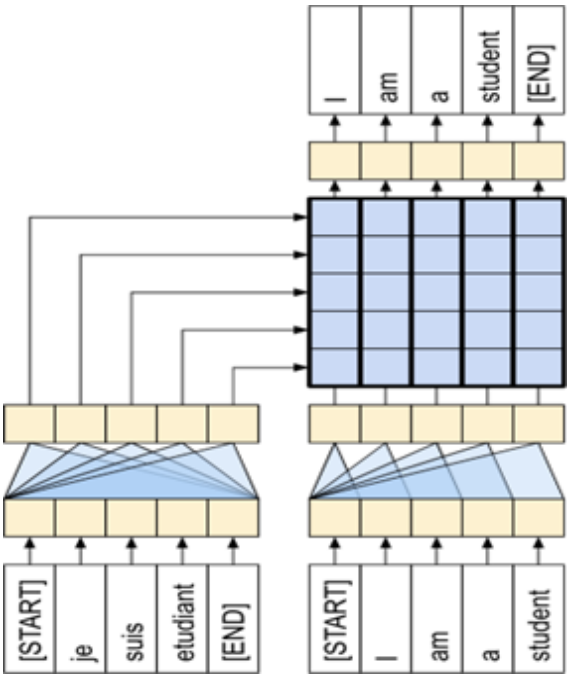

Figure 2: The transformer encoder and decoder

The model will be implemented in three main parts:

1.Input - The token embedding and positional encoding (SeqEmbedding).

2.Decoder - A stack of transformer decoder layers (DecoderLayer) where each contains:

- A causal self-attention later (CausalSelfAttention), where each output location can at-tend to the output so far.
- A cross-attention layer (CrossAttention) where each output location can attend to theinput image.
- A feed-forward network (FeedForward) layer which further processes each output location independently.

1.Output - A multiclass classification over the output vocabulary.

2.Input

The input text has already been split up into tokens and converted to sequences of IDs. Remember that unlike a CNN or RNN the Transformer's attention layers are invariant to the order of the sequence. Without some positional input, it just sees an unordered set not a sequence. So in addition to a simple vector embedding for each token ID, the embedding layer will also include an embedding for each position in the sequence. The SeqEmbedding layer is defined below:

- It looks up the embedding vector for each token.
- It looks up an embedding vector for each sequence location.
- It adds the two together.

3.2.Decoder

The decoder is a standard Transformer-decoder, it contains a stack of Decoder Layers where each contains three sublayers: a Causal Self Attention, a Cross Attention, and a Feedforward. The Feed-Forward layer is below. Remember that layers. Dense layer is applied to the last axis of the input. The input will have a shape of (batch, sequence, or channels), so it automatically applies pointwise across the batch and sequence axes. Next arrange these three layers into a larger Decoder Layer. Each decoder layer applies the three smaller layers in sequence. After each sublayer the shape of $out_seqis(batch, sequence, channels)$.

3.3Output

At minimum the output layer needs layers. Dense layer to generate logit predictions for each tokenat each location. But there are a few other features you can add to make this work a little better:

Handle bad tokens: The model will be generating text. It should never generate a pad, unknown,or start the token (”, ’[UNK]’, ’[START]’). So set the bias for these to a large negative value.
Smart initialization: The default initialization of a dense layer will give a model that initially predicts each token with almost uniform likelihood. The actual token distribution is far from uniform. The optimal value for the initial bias of the output layer is the log of the probability of each token. So include an adept method to count the tokens and set the optimal initial bias.

3.41.Result

As per the result produced by the model, we can say that the model can describe the content of the image in a desired manner. Even though the results are satisfactory but the model is making someerrors when it comes to some complex picks and makes some wrong element analyses. Form the

Figure 3: Result 1

above result we can see that the model was able to predict all the important components from the

Figure 4: Result 2

image and was able to put it in the right context and present it to us. But when it comes to this image the model was not able to predict correctly even though it was able predict the componentsin the image but was not able to put it in right contexts as expected.

3.5 Hypothesis Testing

In order to draw conclusions about a population parameter or probability distribution, statistical reasoning known as hypothesis testing involves analyzing data from a sample, a hypothesis is first formulated in relation to the parameter or distribution. The shorthand for this is the null hypothesis or H0. The null hypothesis is then contrasted with the alternative hypothesis (designated Ha), which is the complete opposite. The hypothesis-testing method decides if H0 can be rejected based on sample data. The alternative hypothesis Ha is valid if H0 is disproved, according to thestatistical result.

For this paper,

Null hypothesis (H_0): Automation of caption generation is not the best way of captioning
Alternative hypothesis (H_a): Automation of caption generation is the best way of captioning

3.6Test Stattistics

There are three tests that can be used to decide whether or not the null hypothesis should be rejected. They are:

1.Chi-squared test
2.T-student test (T-test)
3.Fisher’s Z test

A two-tailed T-student test will be used in this paper.

When comparing the means of two groups that are connected in some way, a t-test is an inferential statistic that assesses whether there is a significant difference.

- Level Of Significance
  The significance level is the likelihood that the null hypothesis will be rejected when it is confirmed (also known as alpha or).
- Level Of Confidence
  The confidence level shows the likelihood that a statistical parameter’s position is correct (suchas the arithmetic mean) measured in a sample survey is also true for the entire population.

Sr.no	Data
1	94.5

2	73.6
3	35.5
4	56.4
5	29.5
6	34.5
7	53.6
8	86.4
Mean	58
Standard Deviation(s)	24.71217167

Level of significance = 0.05 Level of confidence = 95

The number of standard deviations that separate a t-score (or t-value) from the t-mean distribution. The formula to find t-score is:

$t = (x - \mu)/(s/\sqrt{n})$ (1)

where x is the sample mean,

μ is the hypothesized mean,

s is the sample standard deviation,and n is the sample size.

The p-value, also referred to as the probability value, expresses how likely it is that your data occurred under the null hypothesis. Finding the equivalent p-value is possible once we are aware of the value of t. The null hypothesis can be rejected and Automation of caption generation is the best way of captioning if the p-value is less than a certain alpha level (popular choices are .01, .05, and .10).

Calculating t-value:

Step 1: Identify the alternative and null hypotheses.

Null hypothesis (H0): Automation of caption generation is not the best way of captioning.

Alternative hypothesis (Ha): Automation of caption generation is the best way of captioning.

Step 2: Find the test statistic.

The postulated mean value in this situation is taken to be 0.

$t = (x - \mu)/(s/\sqrt{n}) = (58 - 0)/(24.71217167/\sqrt{6})$ $= 6.638 t - value = 6.638$ (2)

Calculating p-value:

Step 3: Calculate the test statistic's p-value.

The p-value is computed using the t-Distribution table with n-1 degrees of freedom. The sample size for this study is n = 8, hence n-1 = 7. It provides a p-value when the observed value is enteredinto the calculator. In this case, the p-value returned is 0.00029366.

We can reject H_0 at the significance level of 0.05 because your p-value does not exceed 0.05. Therefore, we have enough information to conclude Automation of caption generation is the best way of captioning.

4.FUTURE DIRECTIONS FOR RESEARCH

Considering how new the issue of automatically captioning images is, significant advancement has been made as a result of the work of scholars in this area. The effectiveness of image captioning could certainly use some improvement, in our opinion. First, given the rapid advancement of deep neural networks, applying more potent network structures as language and/or visual models will unquestionably boost the effectiveness of image description creation. Second, since captions for photos are simply word sequences whereas images are made up of objects scattered across space,it is crucial to look at the presence and hierarchy of visual concepts in captions. The effective use of the attention mechanism to create picture captions will also remain a key study area since this problem suits the attention mechanism well and because it is proposed to perform a variety of AI-related activities [129]. Third, research on using unsupervised data, such as from photos alone or text alone, to improve image captioning will be promising due to the lack of coupled image-sentence training set. Fourth, present methods generally concentrate on creating general captions describing the contents of images. However, as noted by Johnson et al. [130], picture description needs to be firmly rooted in the aspects of the photographs in order to be relatable to humans and useful in real-life settings. As a result, one of the future study topics will be image captioning founded on image regions. Fifth, while task-specific image captioning is required in some situations, the majority of existing approaches are geared to provide image captioning for generic cases. It will also be fascinating to conduct research on several unique scenarios where image captioning issues arise.

5.CONCLUSION

We categorize picture captioning methods into

many groups based on the strategy used in each method. The strengths and weaknesses of each sort of job are discussed along with representative approaches from each area. We start out by talking about early image captioning research, whichis primarily retrieval- and template-based. Next, neural network-based techniques are the main focus of our research because they produce cutting-edge outcomes. We then separated them into subgroups and examined each subcategory separately because different frameworks are employed in neural network-based methodologies. Following that, benchmark data sets are used to compare state-of-the-art approaches. Finally, we outline potential future avenues for automatic picture captioning research.

REFERENCE

[1] Dumitru Erhan Alexander Toshev, Samy Bengio. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652 – 663, 2017.

[2] Liang Li Jian Yin Anan Liu Zhendong Mao Zhenyu Chen Xingyu Gao Chenggang Yan, Yiming Hao. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):45–51, 2022.

[3] Youssef Mroueh Jarret Ross Tom Sercu Pierre L. Dognin, Igor Melnyk. Adversarial semantic alignment for improved image captioning. *IBM Research, Yorktown Heights, NY*, 2019.

[4] Dmitrij Šešok Raimonda Staniūtė. A systematic literature review on image captioning. *Applied Sciences*, 9(10), 2019.

[5] Yehao Li Zhaofan Qiu Tao Mei Ting Yao, Yingwei Pan. Boosting image captioning with attributes. *ICCV 2017*, 2017.