Image Caption Generation using Transformers

Luka Cvetko
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
lcvetko@nd.edu

ABSTRACT

This paper explores the usage of Transformers in Automatic Image Caption generation, focusing on enhancing Content Image Based Retrieval Systems(CBIR). A novel approach is presented that combines the strength of a Convolutional Neural Network with the semantic understanding of Transformers. Using the Flickr8k dataset, the data was preprocessed to create unique vocabulary from the caption list. The InceptionV3 model is utilized to extract image features, followed by caption generation from the Transformer model. BLEU score was implemented to calculate the accuracy of generated captions. The results indicate that the use of Transformers can generate coherent and appropriate captions with complex relationships between different object from an image. Despite limitations such as a smaller dataset, and small number of training epochs, our findings indicate that Transformers can offer a substantial improvement over traditional CBIR methods.

I. GOAL

In today's realm of digital information, Content-Based Image Retrieval (CBIR)[3] systems have become vital in the process of extracting images from databases. These systems enable extraction of mages based on a query request to a large database using features such as shape, texture, and color. One of the main challenges in advancing such systems is to develop a method in which the system can understand and interpret a request a way a human would. Automatic image captioning aims to overcome the limitations of CBIR by providing a descriptive and semantic understanding of the image contents in, which aligns more closely with the human perception in querying methods. While conventional methods are effective matching the visual features, they often fall short when the entire image context is considered, leading to more broader query results instead of fixating solely on a small part of images. Typical image recognition systems using a CBIR approach, where given a prompt such as "tree" the user will be given multiple instances of images that represent a tree in the database. One of the main advantages that automatic image captioning has over CBIR is that it allows the user much more choice and freedom when searching different queries. This gap highlights a necessity to develop a more nuanced approach to image analysis and retrieval. Machine-learning algorithms more have provided the needed tools to bridge this gap. The current best practice in the implementation of these systems is the use of a Recurrent Neural Networks (RNN) [6] which will generate the caption and use a Convolutional Neural

Network (CNN) [7] for the image classification. One of the best(RNNs for this task is LSTM(Long Short-Term Memory)[8] which can tackle sequence-based prediction problems, such as Google search. LSTM models process data linearly and sequentially, which can be a limitation with complex dependencies and bigger datasets. With these limitation in mind, the transformer model, introduced in the paper "Attention is All You Need" by Vaswani et al.[9] was created. Compared to other sequential data processing models, transformer model employs a self-attention mechanism that allows them to weight and integrate information from different parts of the input in parallel. This attribute is crucial for tasks that require comprehensive understanding of the context such as generation of descriptive captions for images, and it enables transformers to understand the context of the entire input data more effectively than an LSTSM. This paper introduces a concept to create an automatic image caption generator based on a transformer model, enhancing the capabilities of the query search. Presenting a hybrid architecture approach that combines the strength of CNNs to extract image features and use the contextual power of transformers to generate an appropriate caption for said image. The goal of this project is to create a system that can correctly identify and caption an image with proper relations between the objects present in the image.

The following is the user interaction with the system highlighting the input and output process and explain how the user will interact with the end-system.

1. User uploads an image of their choice to be captioned
2. Program will take that image and perform preprocessing on the image and then feed it through a CNN where the images feature vectors are extracted.
3. The feature vectors are fed as input into the Transformer model which will generate the caption.
4. The final output provided is the original image with generated caption describing the image underneath.

II. METHOD

II.I Dataset

The dataset that was used is the Flickr8k dataset[5]. This dataset contains around 8000 images that are each paired with five different captions that describe the image. The data will be preprocessed. First an extraction of all the unique words will be done to extract a unique vocabulary out of the caption data. Next, the vocabulary will be cleaned, meaning that all punctuations will be removes, single characters will be removed, and all numbers will be removed. After the cleaning method is completed, all the captions and images will be saved to two different list. This enables to load images using the set path, so all images will be labeled the same as all the captions, so that one can locate them in their separate list afterwards. Additional tags will be added at the start of each caption with the <start> tag and at the end of each caption with the <end> tag to indicate to the transformer model where the start and where the end of the caption is located. Next the dataset is split into a training and validation split using an 80/20 split, where 80% of the data is used for training and 20% is used for testing and validation.

II.II Image feature extraction using InceptionV3

After the dataset has been determined and split, the first part of the hybrid approach to image caption generation will be conducted. A CNN model, specifically the InceptionV3[11] model will be used to extract feature vectors from the image datasets. InceptionV3 is characterized by its architecture, which includes multiple convolutional filters within the same layer. It is trained on a large dataset called the ImageNet dataset with varied images of objects in different environments with different interactions between them. With this design InceptionV3 can capture a wide range of feature from an image such as edges, textures, complex patterns with varied levels of abstraction. What makes it stand out, is the model's depth and width which are carefully balanced in order to optimize its performance, making it a great tool in identifying intricate details of an image. The model utilizes techniques to reduce dimension sizes, such as factorized convolutions and aggressive regularization techniques, which reduce computational needs without compromising the quality of extracted features from an image. This in turn provides the transformer model with a comprehensive list of image features which are essential to generate an accurate and detailed image caption. After the features are extracted, they are saved into a .npy file format which is later used as input for the transformer model.

II.III Transformer model to extract captions

Following feature extraction process from InceptionV3, the features get input into the Transformer model. The main architecture can bee seen in Figure 1.

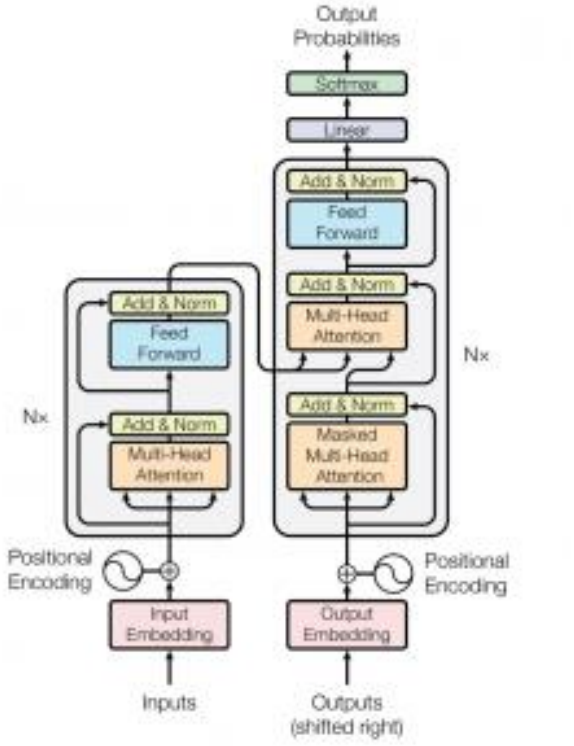


Figure 1. Transformer model architecture

The main structure of the transformer model is based on a layered structure consisting of an encoder and a decoder. The encoder processes the input which are the extracted image features, while the decoder generates the output which are the generated captions. Both components possess two crucial subcomponents the Muti-Head Self Attention layer and a fully connected feed-forward network. The process of which an image caption is generated follows this path and will go through these

components. First, the extracted features will be normalized to feature vectors, then positional encoding will be added which are used to maintain spatial information, since the transformer model does not process the data in a sequential manner. After this, the features with their encoding are inputted into the Encoder and the Multi-Head attention layer, this mechanism allows the model to weight different parts of the image, focusing on feature that could be relevant to caption generation. The multi-head aspect of the mechanism involves running several of these attention processes in parallel , this in turn leads to a more nuanced and comprehensive understanding of the image features. Next the output from the Encoder is inputted into the Decoder where a similar process takes place, but in this case the attention mechanism focuses on the decoder output, ensuring that each word generated in the caption considers the entire image context provided by the Encoder. Finally, the Decoder generates the caption one word at a time, using both the previous work and the output from the Encoder. Then a prediction is made on the next word based on the current state and the addition of all the attention weighted features. This continues until the <end> token is reached, which signals the end of the caption and its completion.

II.IV Loss Function for model optimization

In order for the model to learn with each generation a proper loss function needs to be defined. The approach in this paper, defines a custom loss function based on Cross-Entropy. The main goal of this loss function is to measure the difference between the predicted caption and the actual (ground-truth) caption. First the loss function creates a mask to identify any padding added and excludes them from the calculation. Next, Cross-Entropy is used to measure the difference between the probability distribution of predicted captions and actual captions. This method is incorporated into the training loop. Using TensorFlow automatic differentiation capabilities, gradients of the loss function are calculated and update the model parameters via backpropagation. Two metrics are used, mean loss and sparse categorical accuracy. The mean loss provides an average of the loss over the training batches, offering insight into the overall convergence of the model, while the accuracy metric gives a direct measure of how often the model correctly predicts the next word in the caption. The model is then trained on 30 epochs.

III. EXPERIMENTS AND RESULTS

III.I Implemented Transformer model architecture

The model transformer mode architecture used in our implementation can be observed in Figure 2.

```
layer_num = 4
dim_mod = 512
ff_dim = 2048 #feed forward dimension
heads = 8
row = 8
col = 8
target_vocab_size = top_k + 1
drop_rate = 0.1
```

Figure 2. Transformer model architecture implemented

The model comprises of 4 layers in the Encoder and Decoder. The dimensionality of the vectors is set at 512, which offers a good compromise between amount of information acquired and computational usage. The feed-forward network in each transformer block has a dimension of 2048 indicated the capture of complex relationships. An 8-head multi-attention mechanism is used which allows for a plethora of parallel processing. The input data is structured as an 8x8 grid, and the vocabulary size accommodates a broad range of tokens with the addition of the end of sequence token. A dropout rate of 0.1 is implemented for regularization to prevent overfitting.

III.II Evaluation of generated caption quality with BLEU score

To determine the efficiency of the model generated captions, Bilingual Evaluation Understudy (BLEU)[10] was used. The BLEU score quantifies how similar the generated caption is to the ground truth caption. This score is calculated by the n-gram overlap between the generated caption and ground truth caption. N-grams in this case are sequences of words from a given sample of text. The 1-gram, 2-gram,3-gram and 4-gram BLEU mode was used to calculate the model efficiency taking the highest in each category. The general pipeline of evaluation follows the order. First the captions will be generated by the model, next the reference captions for each image in the test set will be preprocessed by doing text modifications. Finally, the BLEU score will be calculated by comparing the generated and ground-truth captions.

III.III Baseline method

To establish a comparative analysis for our model a baseline method had to be defined. The implementation was a method that relies solely on the CNN for generating captions. This will serve as a fundamental benchmark to

which the transformer model is compared to. This method uses a CNN to extract image features from an image in this case the ResNet50 and convert them into a comprehensive list of items detected in an image without any relationships between the detected objects. Once the CNN extract the features a concatenation method is used to combine all detected features and calculate the BLEU score when compared to the ground truth caption. Figure 3. demonstrates the baseline method generated captions and BLEU score.

Figure 3. baseline method generated caption and BLEU score

The baseline method generated a non-relationship-based caption with a BLEU score of 7.45. This indicates some overlap in the words with some of the captions such as the word motor and bike, but no relationship can be made between the objects in the image leading to a small overlap and a low BLEU score.

III.IV Transformer model results

The Transformer model achieved significantly higher BLEU scores compared to the baseline CNN method. In the same image that was used as the baseline the Transformer model achieved a much higher BLEU score. This was particularly noteworthy, indicating that the Transformer model is much more superior in constructing more coherent and contextually appropriate captions.

The detailed BLEU score results are, and generated captions can be seen below:

- Transformer Model BLEU score: 25.0
- Baseline Model BLEU score: 7.45
- Actual caption: dirt bikers on trail
- Transformer caption: two people on motorbikes
- Baseline caption: mountain bike crash helmet alp motor scooter moped

In terms of accuracy and precision, the Transformer model demonstrated a higher ability to select relevant words and phrases that accurately describe the images. Contrasted with the baseline model, which, while effective in identifying key objects, often lacked in providing detailed descriptions that took relationship of objects in the image into account. More detailed images of generated captions for different images can be seen under the Appendix section.

IV. CONCLUSION

This paper attempted to explore the efficiency of Transformers on Automatic Image Caption Generation with particular interest in enhancing Content-Based Image Retrieval Systems (CBIR). With a comparative analysis to the baseline CNN method, the research demonstrates that the Transformer model has superior capability and can generate more intricate and detailed captions. The main findings indicate that the Transformer model outperforms the baseline CNN approach in the evaluation of BLEU score. The Transformer ability to understand greater context of relationship between objects in an image were much more descriptive and aligned with human-like perception and language. One key finding was that there seems to be a correlation between lower BLEU score and image complexity. This can suggest that while the transformer model excels in different context, its performance can be hindered by the complexity of the visual data. This leads us to future improvements that could improve upon the current model. First training dataset only contained 8000 images, using a bigger dataset such a Microsoft Common Objects in Context (COCO)[4] which contains more than 12,000 images should be used for future implementations. Small number of epochs was also used in this study only 30, a recommendation of at least 50 epochs should be used for future research. Despite these areas of improvement, the Transformer model demonstrates a notable improvement in general CBIR methods, particularly in image understanding and retrieval. The entire project including the Transformer and baseline methods can be found on the following GitHub link: https://github.com/Windtwist/nlp-proj

V.REFERENCES

[1] Al-Malla, M.A., Jafar, A. & Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20 (2022). https://doi.org/10.1186/s40537-022-00571-w

[2] Wang, C., Zhou, Z., & Xu, L. (2021). An integrative review of image captioning research. Journal of Physics: Conference Series, 1748(4), 042060. https://doi.org/10.1088/1742-6596/1748/4/042060

[3] What Is Content-Based Image Retrieval? | Baeldung on Computer Science - https://www.baeldung.com/cs/cbir-tbir

[4] COCO Dataset | Papers With Code - https://paperswithcode.com/dataset/coco

[5] Flickr30k Dataset | Papers With Code - https://paperswithcode.com/dataset/flickr30k

[6] What Are Recurrent Neural Networks? | IBM. www.ibm.com/topics/recurrent-neural-networks.

[7] What Are Convolutional Neural Networks? | IBM. www.ibm.com/topics/convolutional-neural-networks.

[8] Brownlee, Jason. "A Gentle Introduction to Long Short-Term Memory Networks by the Experts." MachineLearningMastery.com, 6 July 2021, machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[10] "Evaluating Models." Google Cloud, cloud.google.com/translate/automl/docs/evaluate.

[11] C. Wang et al., "Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model," in IEEE Access, vol. 7, pp. 146533-146541, 2019, doi: 10.1109/ACCESS.2019.2946000.

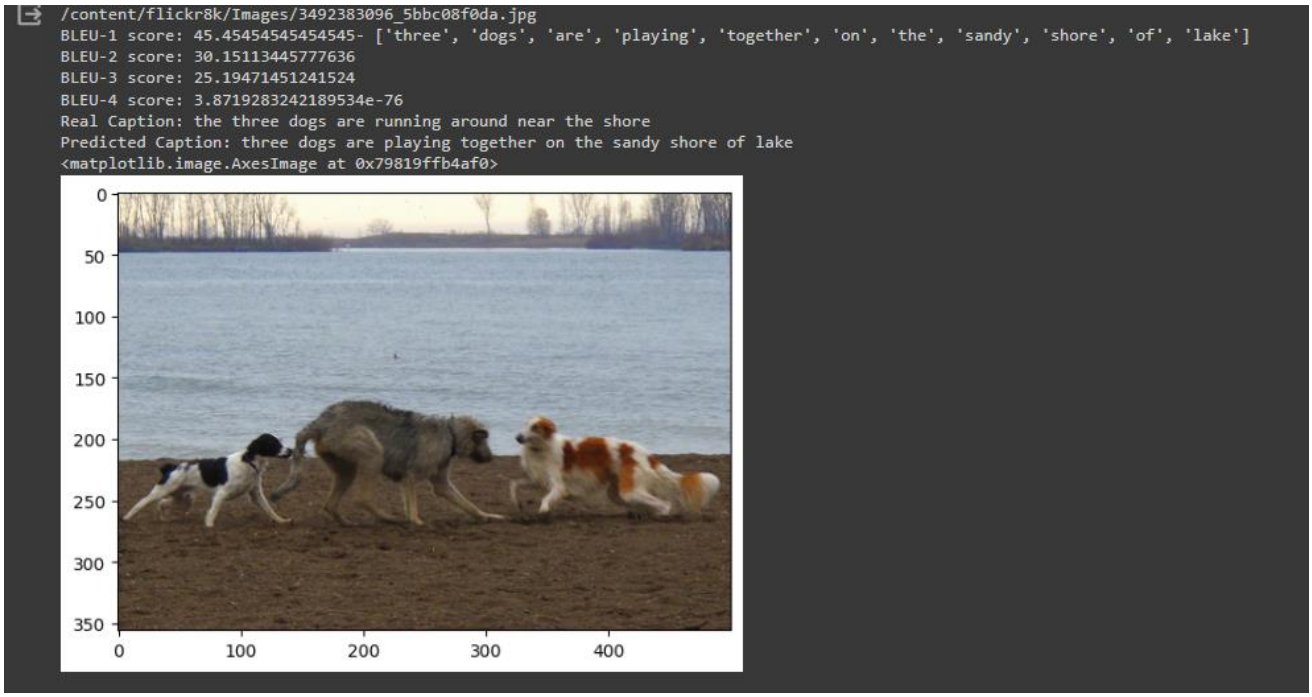
[12] Jia, Xu. Guiding the Long-Short Term Memory Model for Image Caption Generation. 2015, openaccess.thecvf.com/content_iccv_2015/html/Jia_Guiding_the_Long-Short_ICCV_2015_paper.html.

[13] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

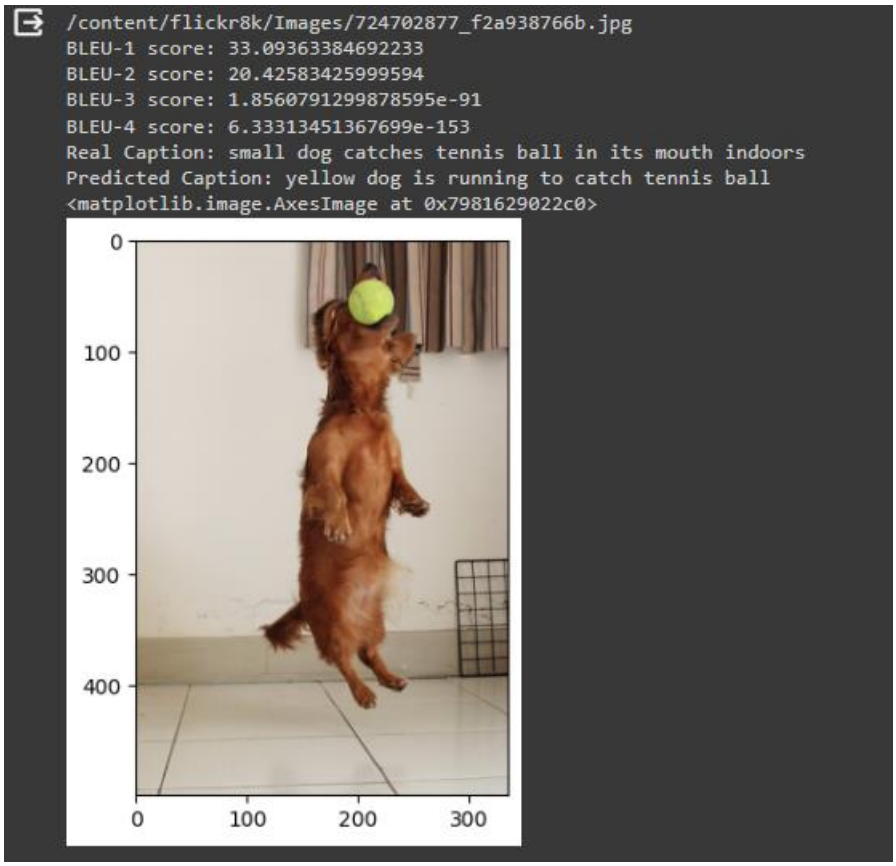
[14] Shen, X., Liu, B., Zhou, Y. et al. Remote sensing image caption generation via transformer and reinforcement learning. Multimed Tools Appl 79, 26661–26682 (2020). https://doi.org/10.1007/s11042-020-09294-7

[15] R. Castro, I. Pineda, W. Lim and M. E. Morocho-Cayamcela, "Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks," in IEEE Access, vol. 10, pp. 33679-33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

VI. APPENDIX



Appendix I. BLEU score: 45, generated caption: three dogs are playing together on the sandy shore of lake, actual caption: the three dogs are running around near the shore



Appendix II. BLEU score: 33.09, Generated caption: yellow dog is running to catch tennis ball, Real Caption: small dog catches tennis ball in its mouth indoors

Appendix III. BLEU score: 45.45, Generated caption: woman in colorful shirt is taking picture from the water fountain, Real Caption: woman with green shirt takes drink from water fountain

```
/content/flickr8k/Images/3104400277_1524e4f758.jpg
BLEU-1 score: 25.0
BLEU-2 score: 7.458340731200295e-153
BLEU-3 score: 1.6896185209462902e-183
BLEU-4 score: 1.2882297539194153e-229
Real Caption: dirt bikers on trail
Predicted Caption: two people on motorbikes
<matplotlib.image.AxesImage at 0x7981505bf2b0>
```

Appendix IV. BLEU score: 25.0, Generated caption: two people on motorbikes, Real Caption: dirt bikers on trail