Generating image captions using Encoder-Decoder architecture.

Mohammad Adil Shaikh
BTech Artificial Intelligence
Mukesh Patel School of Technology Management & Engineering (NMIMS),
Mumbai, India.
mohammadadil.shaikh083@nmims.edu.in

Mohammed Az Syed
BTech Artificial Intelligence
Mukesh Patel School of Technology Management & Engineering (NMIMS),
Mumbai, India
mohammedaz.syed24@nmims.edu.in

Jash Vasa
BTech Artificial Intelligence
Mukesh Patel School of Technology Management & Engineering (NMIMS),
Mumbai, India
jash.vasa011@nmims.edu.in

Chetan Yadav
BTech Artificial Intelligence
Mukesh Patel School of Technology Management & Engineering (NMIMS),
Mumbai, India
chetan.yadav090@nmims.edu.in

Abstract: This research explores the exciting field of image captioning, where models can generate descriptions for images through a combination of deep learning and natural language processing. We trace the development of early rule-based systems through the significant influence of recurrent neural networks (RNN) and convolutional neural networks (CNN). This article discusses current issues regarding the models for image captioning and has tried two approaches to the problem. In addition to the technical details, we cover many graphics applications and dive into how they can revolutionize automation, human-computer interaction, and accessible content.

Keywords: CNN, RNN, Captions, LSTM, Encoder-Decoder architecture, Flickr8k, VGG16, Xception, InceptionV3.

I. Introduction

The combination of Deep learning (DL) and natural language processing (NLP) is changing the way people interact with and interpret information found in modern artificial intelligence (AI). A good example of this collaboration is the process of creating the correct look and content of the images. Thanks to the power of deep learning models, the project called "Image Captions" has achieved unprecedented development. Images contain a lot of information because they are an important part of the human experience. However, pixel-to-word connectivity has proven to be an ongoing challenge for robots to analyse images well and convey their meaning in words. This technology is now required in many areas, from autonomous navigation and human-machine communication to content suggestion and accessibility.

The task of image captioning poses a distinct multidisciplinary problem, requiring the seamless fusion of deep learning techniques for comprehending visual content and natural language processing models for producing logical, contextually rich captions. The proliferation of deep learning architectures, especially convolutional neural networks (CNNs) for image analysis and recurrent neural networks (RNNs) or transformer-based models for natural language generation, has enabled the shift from a pixel-centric understanding to a semantically meaningful representation of images.

This research paper explores the state-of-the-art approaches for combining deep learning and natural language processing to generate image captions. It explores the underlying methodologies, structures, and obstacles that this intriguing endeavour faces. We hope to present a thorough overview of the field, provide light on possible uses, and spur additional progress in the effort to equip machines with the capacity to detect and describe the visual environment by clarifying the current state of the art.

We will also investigate possible uses, such as enhanced accessibility of content, interaction between humans and computers, and creative solutions in fields like healthcare, autonomous systems, and more.

II. Literature Review

Image captioning, the task of automatically generating textual descriptions for images, represents a convergence of computer vision and natural language processing (NLP) and has seen remarkable progress in recent years. This section provides an overview of key developments, methodologies, and challenges in the field of image captioning, highlighting the contributions of deep learning techniques.

Early Approaches to Image Captioning

Automatic generation of descriptions of images or graphics earlier used to depend heavily on computer vision techniques and rule-based methods. Some of the early approaches depended on Handcrafted rules, hand-annotated datasets, and Template-based approaches.

Deep Learning Revolution

With the development of deep learning, image captions have also undergone a major change. Deep neural networks, specifically Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) enable end-to-end learning from data. Convolutional neural networks are a combination of CNNs for image extraction and RNNs for sequence generation. Notable contributions in this research are Vinyals et al. [1] first proposed the idea of using CNNs to encode images and generate captions using LSTM (long-term memory) networks. This is a revolution in image understanding, from self-generated models to deep learning, Ren et al.'s [2] research popularized the idea of using local CNNs to identify objects in images. Thanks to this technology, image models can detect and define specific areas in images and Vaswani et al. [3], introduced the Transformer architecture, which significantly impacted the field of natural language processing (NLP) and had important implications for image captioning tasks.

Datasets and Evaluation Metrics

This field has become successful thanks to the emergence of large image repositories such as MS COCO [8] and Flickr30k [9]. This information makes it easier to train and evaluate the model. Despite their shortcomings, criteria such as BLEU, METEOR and CIDER have become industry standards that determine the quality of labels.

Applications and Beyond

Descriptions are used in many areas, such as making websites more accessible for the visually impaired, improving content recognition, helping driverless vehicles understand their environment, and supporting human-machine interaction.

In summary, the development of deep learning and NLP technology has led to significant advances in image collection. Depending on the area of development, innovations such as supportive learning and multimodal learning can improve the image.

III. Proposed Model

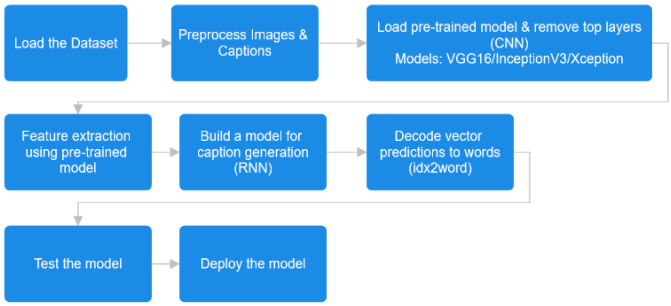

Fig 1. Architecture for Encoder-Decoder model

Dataset Used

We have used the flickr8k dataset [5], which consists of 8091 images and 5 captions for each image. The images are collected from the Flickr photo-sharing website. The captions in this dataset are human annotated, hence more natural, and fluent as compared to automatically generated captions. This dataset is widely used in the field of image captioning as the diverse dataset covers a wide range of objects and activities.

Fig 2. Sample image from the dataset.

Captions associated with the image:

1. A soccer game at sunset
2. A soccer game is being played as the sun sets.
3. Green sports field with players beneath a blue sky with pink clouds
4. People playing soccer on a soccer field during sunset
5. Soccer players on a field from a distance.

Transfer learning

Transfer learning [11] is an ML technique, where a model is trained on one task and is utilised for a different task. The previous weights and biases are leveraged for the new application (in our case feature extraction). This technique has immensely helped in increasing the efficiency and results of the Encoder-Decoder Architecture used in various applications.

First approach

In this approach we have opted for the encoder-decoder architecture [4], where we use CNN as the encoder and RNN as the decoder, utilising CNN for image feature extraction and RNN specifically LSTM for caption sequence generation.

We have utilised the pre-trained model VGG16 [6] as our CNN, having removed the final layers, one can utilise this model efficiently for feature extraction of an image. VGG16 is a straightforward architecture which consists of 16 weight layers. It is pre-trained on the ImageNet dataset due to which it is an excellent choice for transfer learning. The model has great flexibility as it can handle input images of various sizes without making drastic modifications.

RNNs are a type of neural network that is well suited when we are dealing with sequential data. This is because they maintain a hidden state that helps retain information from previous time steps. The output of a CNN (in our case the VGG16 model) is passed to an RNN in our encoder-decoder architecture, the output is in the form of a fixed-length vector that has features or visual information about the image encoded in it. While RNNs are the natural choice for sequential data, they tend to face the vanishing gradient problem. To overcome this drawback of RNN, we use the LSTM network as it can capture long-term dependencies. In our application of image captioning, it is required to understand the relation between objects and actions in an image, this can span multiple time steps and lead to long-term dependencies. Hence, applying a vanilla RNN might lead to improper handling of long-term dependencies. Therefore, LSTM networks are better suited for this application. The text sequences or the captions will be of varied lengths across the dataset, and LSTMs will be better at handling variable-length sequences.

Second approach

In this approach, we have taken the same approach for the architecture as above (*Encoder-Decoder architecture*). Here, instead of utilising the VGG 16 model for feature extraction, we have opted for the deep convolutional model developed by Google, Inception V3 [7].

Inception V3 is characterized by its deep architecture with 48 layers and it has also been trained on the ImageNet dataset. This model yields high accuracy on different image classification tasks. The main innovation in this pre-trained model is the use of Inception modules, these are the building blocks that allow the network to extract image features at different scaled and abstraction levels. The Inception modules use a combination of different kernel sizes to capture features at different spatial resolutions. With the help of the convolutional layers of the model, we can extract meaningful features to pass to other models.

Third approach

In this approach also, we have taken the same approach for the architecture as above (*Encoder-Decoder architecture*). Here, we have opted for the deep convolutional model developed by Google, Xception.

Xception, which stands for “extreme inception,” is a deep convolutional neural network designed for image classification. This network was introduced by Francois Chollet in the year 2017, who works at Google, Inc. and is also the creator of Keras1.
The model is 71 layers deep. It relies solely on depthwise separable convolution layers. It splits the convolution into two separate steps, i.e., depthwise convolution and pointwise convolution. Xception takes a more extreme approach to depthwise separable convolutions, leading to a deeper and more parameter-efficient network designed to capture fine-grained image features. Thus, making Xception an extreme version of the Inception Model.

Metric

In the intricate field of efficacy evaluation of our implemented model, we have chosen the BLEU metric, a well-established measure renowned for its effectiveness in assessing the efficiency of machine translations. The essence of BLEU lies in its meticulous calculation, wherein it scrutinizes the alignment of n-grams—contiguous sequences of n-items, usually words—between the candidate and reference translation

The process unfolds with a comparative analysis of how well the n-grams translation aligns with the reference translation. Notably, this assessment is nuanced by a position-independent consideration, meaning that the metric considers the overall occurrence of matches rather than their specific order. In essence, it gracefully accommodates variations in word order and structure, making it a robust metric for evaluating the fluency and accuracy of translations.

The careful counting of matching n-grams is the key component of BLEU's computation. The degree of alignment between the candidate and reference

translations is indicated more by a greater count. Because of its position-independent matching process, BLEU can accurately and nuancedly capture the essence of translation quality, considering the characterisation of semantics and language structure.

Consequently, the BLEU metric provides a quantitative measure that is closely correlated with the fidelity and accuracy of the model's output, making it a trustworthy benchmark for assessing the calibre of potential translations. Essentially, it offers a useful and complex prism by which to evaluate the model's ability to produce translations that accurately capture the complexities and subtleties of the original language.

IV. Results

Fig 3. Image from Test Set.

Predicted Caption:

Child is sliding down slide.

Actual Captions:

1. A boy goes down an inflatable slide.
2. A boy in red slides down an inflatable ride.
3. A boy is sliding down in a red shirt.
4. A child going down an inflatable slide.
5. A young boy sliding down an inflatable is looking off camera.

Model (Feature extraction)	BLEU Score
VGG16	0.67
Inception V3	0.59
Xception	0.56

Table 1. Performance of approaches on the Flickr8k dataset

V. Conclusion

Image captions have many important applications, including human-computer interaction, content recognition, accessible content, and more. The ability to translate the content of the image to its respective textual description has had an impact in many areas, improving communication, accessibility, and decision-making.

The development of end-to-end learning is important for the transfer of world views and texts, as images can now be understood and explained without the need for special rules or rules of speech. Contributions from models such as Tell and Tell, Quick R-CNN, and tracking techniques have advanced the field, allowing for a deeper understanding of visual content and the betterment of natural language. Big data sources like MS COCO [8] and Flickr30k [9] are now available; It turns images and their respective captions into recorded research and developmental tools, allowing models to learn from more detailed and varied information.

There are still problems existing with blurry images, longevity of text, and comprehension of unclear images. More research and creativity is needed to solve these problems. As artificial intelligence develops and expands its utility, image captions can improve human-computer interaction and open new opportunities for research and development.

In conclusion, Image Captioning demonstrates the perfect collaboration between natural language processing and deep learning. We hope that this research, which allows machines to see, understand and explain the worldview, will be a useful tool.

References

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and Tell: A Neural Image Caption Generator. *ArXiv*. /abs/1411.4555

[2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *ArXiv*. /abs/1506.01497

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*. /abs/1706.03762

[4] S. Sehgal, J. Sharma and N. Chaudhary, "Generating Image Captions based on Deep Learning and Natural language Processing," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 165-169, doi: 10.1109/ICRITO48877.2020.9197977.

[5] S. C. Gupta, N. R. Singh, T. Sharma, A. Tyagi and R. Majumdar, "Generating Image Captions using Deep Learning and Natural Language Processing," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-4, doi: 10.1109/ICRITO51393.2021.9596486.

[6] Tammina, Srikanth. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications (IJSRP). 9. p9420. 10.29322/IJSRP.9.10.2019.p9420.

[7] S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty and S. Saha, "Image Captioning Using Inception V3 Transfer Learning Model," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1103-1108, doi: 10.1109/ICCES51350.2021.9489111.

[8] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common Objects in Context. *ArXiv*. /abs/1405.0312

[9] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *ArXiv*. /abs/1505.04870

[10] Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.

[11] Hussain, Mahbub & Bird, Jordan & Faria, Diego. (2018). A Study on CNN Transfer Learning for Image Classification.