Image Captioning Tool

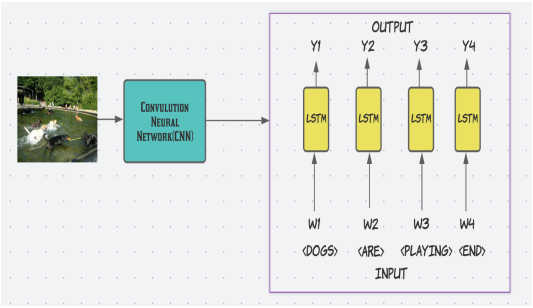
Dinesh Mannari
Indiana Univeristy Bloomington
dmannari@iu.edu

Abstract

Image caption has increasingly caught the attention of many researchers in the domain of artificial intelligence owing to the quick growth of AI in recent years and has grown to be a challenging task.A key component of scene understanding,which integrates the expertise of machine learning and natural language understanding,is image caption,which automatically generates natural language descriptions in accordance with the information observed in an image. The use of image captions is widespread and important,for instance in the execution of human-computer interaction.The attention mechanism,which is a fundamental component of computer vision and has lately become widely used to challenges involving the creation of image captions,is the subject of this paper's overview of related techniques.The advantages and pitfalls of these methodologies are also discussed,along with the most prominent data set and assessment criteria in this area.This research concludes by highlighting several unresolved issues with the image caption task.

1 Introduction

With the ease and affordability with which visual data, such as pictures and films, can now be acquired, a wealth of information for solving practical issues is now available. Due to the abundance of data, there is a desire for automatic visual comprehension and content summarization, that cannot be done in real time. Models won't inherit this flexibility until a suitable hybrid architectural combination is found, as humans can understand themselves without a description of a picture. As a result, employing various deep learning models, image captioning is utilized to explain images. Utilizing computer vision and natural language processing, captioning is a method for explaining an image's content In order to produce human-readable words, captioning enables the model to recognize not only the items in an image but also their relationships with one another. The CNN-RNN architecture, which uses Neural Network layers for feature extraction on input data and RNN to make predictions based on time series analysis, is used to generate captions. Knowledge of Python programming, Keras and its various image recognition functions, working of Convolution neural network, pre-trained model, and transfer learning are all required. I looked through a few datasets on the internet, including Flickr 30k, MS COCO, COCO Captions,Textcap, and Flickr 8k. I used the Flickr 8k dataset to train and test the model. We can use the other dataset if it is more convenient for us. The main difference is the size of the dataset. The diagram below outlines how we will use the input and produce the intended results.



2 Literature Survey

We discuss the three major categories of existing image captioning methods in this section: template-based image captioning, retrieval-based image captioning, and novel caption generation. To generate captions, template-based techniques use fixed templates with blank slots. The multiple elements, actions, and attributes are first recog-

nised in these systems, and then the gaps in the templates are filled. Farhadi et al. [1], for example, use three different elements of a scene to fill the template slots for generating image captions. Kulkarni et al. [2] use a Conditional Random Field (CRF) to detect objects, attributes, and prepositions before filling in the blanks. Template-based approaches can produce grammatically correct captions, but because the templates are predefined, they cannot produce variable-length captions. Visual and multimodal spaces can be used to generate new captions. The visual content of the image is first analysed in these systems, and then captions are generated from the visual content using a language model [3], [4]. These methods can produce new, more semantically accurate captions for each image. Deep machine learning is used in the majority of novel caption generation techniques. As a result, the primary focus of this paper is on novel image caption generation methods based on deep learning. Deep learning-based image captioning methods can also be divided into three categories based on their learning techniques: supervised learning, reinforcement learning, and unsupervised learning. Other Deep Learning includes reinforcement learning and unsupervised learning. Captions are typically generated for an entire scene in an image. Captions can, however, be generated for different areas of an image, as in Dense captioning. Captioning methods for images can use either a simple Encoder-Decoder architecture or a compositional architecture. Data labeling is frequently difficult. As a result, the recent emphasis has been on reinforcement learning and unsupervised learning-based image captioning techniques. A reinforcement learning model chooses an action, receives a reward signal, and changes states. The model attempts to choose the action with the greatest long-term reward. To provide the expectations of a value function, continuous state and action information is required. Modern reinforcement learning techniques have many limitations, such as the lack of value function guarantees and non-specific state-action information. Policy gradient methods [53] are a type of reinforcement learning that uses gradient descent and optimization techniques to select a specific policy for a specific action.

3 Implementation

A captioning model relies on two components, A CNN and an RNN. Captioning is all about merging the two to combine their most powerful attributes i.e spatial information and recognize objects in image. RNNs work well with any kind sequential data, such as generating a sequence of words. So by merging the two, we can get a model that can find patterns and images and then use that information to help generate a description of those images.

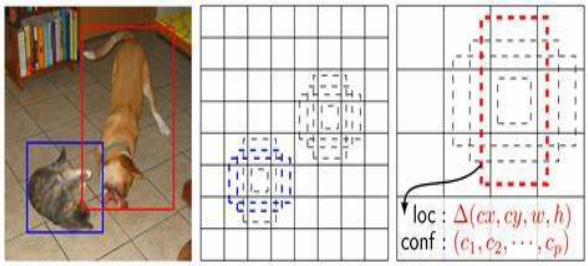
3.1 Approach

If we are given a images as above to write a caption which describes it: Based on how these objects are placed in the image and their relationship to each other. You might think that the kid is posing near water or the kid is playing in the beach. After collecting these visual observations, you could put together a phrase that describes the image as “A happy kid playing in the beach”.

3.2 Challenges

Dual priorities: classification and localization of objects, The first major complication of object detection is its additional goal: we want to classify image objects as well as determine their positions, which is known as the object localization task. To address this problem, researchers frequently employ a multi-task loss function that penalizes both misclassifications and localization errors. Fast R-CNN improves accuracy while dramatically increasing speed because the classification and localization tasks

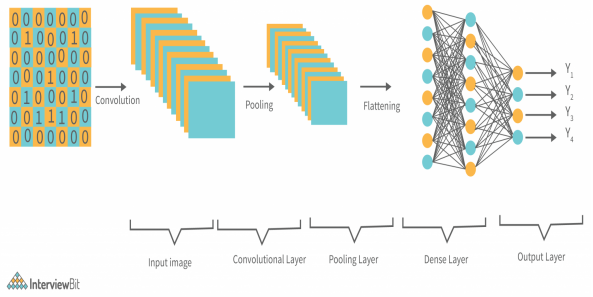
are optimized using a single unified multi-task loss function. Each candidate region that may contain an object is compared to the true objects in the image. Another significant barrier is the scarcity of annotated data currently available for object detection. Object detection datasets typically include ground truth examples for a dozen to a hundred different object classes, whereas image classification datasets can include up to 100,000 classes. Furthermore, crowdsourcing frequently results in free image classification tags (for example, by parsing the text of user-provided photo captions). Obtaining ground truth labels and accurate bounding boxes for object detection, on the other hand, is extremely time-consuming.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

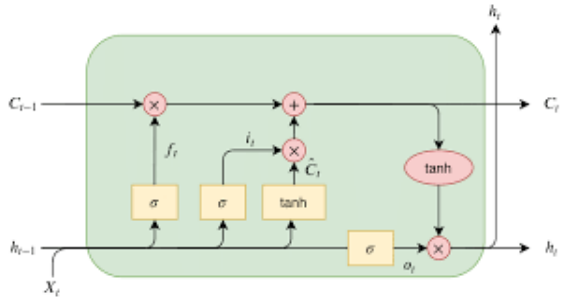
3.3 CNN for Image extraction

Image captioning techniques are broadly classified into two types: those that use the template method and those that use the encoder decoder structure. A CNN, or Convolutional Neural Network, is a deep learning neural network designed specifically for processing structured arrays of data such as images. Convolutional neural networks excel at recognizing key features and patterns in input images such as lines, circles, and even eyes and faces. Another feature that makes CNN so powerful is its ability to work directly on raw images without pre-processing. There are numerous applications for it, including photo and video recognition, image classification, medical image analysis, computer vision, natural language processing (NLP), and so on. Convolution is a type of linear operation in which two functions are multiplied to produce a third function that expresses how the shape of one function is modified by the other, as denoted by the word "Convolution" in Convolution Neural Network. A feed-forward neural network with up to 20 or 30 layers is known as a convolutional neural network. Handwritten digits can be recognized with three or four convolution layers, and human faces can be distinguished with 25 layers. Convolution Layer, Pooling Layer, and Fully Connected Layer are the three basic layers of a CNN architecture. In addition to these three layers, the dropout layer and the activation function are important parameters.



3.4 LSTM For producing captions

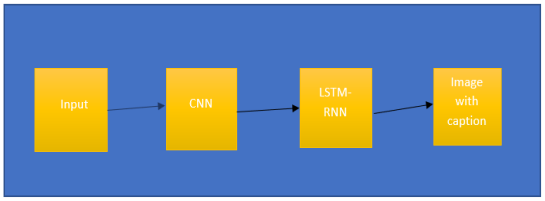
The main disadvantage of RNN was that due to back propagation, vanishing/exploding gradient effects could occur if the sequence was very long or if the neural network had more than one hidden layer. Long Short Term Memory (LSTM) was created to address these issues. LSTM is a type of RNN architecture that addresses vanishing/exploding gradients and allows for long-term dependency learning. With cutting-edge performance in speech recognition, language modelling, translation, and image captioning, LSTM has risen to prominence. When compared to RNN, LSTM can keep information for longer periods of time. It primarily employs long-term memories (information collected a long time ago) and short-term memories (information collected a few timestamps ago) in conjunction with the current event to generate a new modified long-term memory.



3.5 Proposed Model

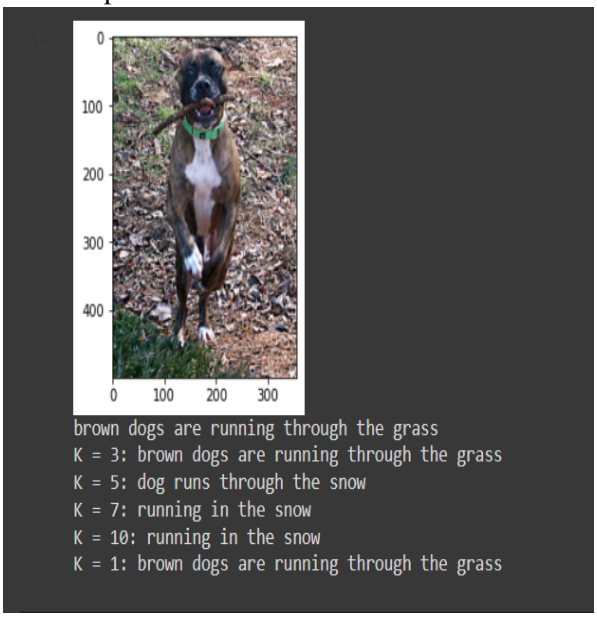
Deep learning is used in our model for image captioning. For image classification, we primarily employ two techniques: CNN and LSTM. So, to create our image caption generator model, we will combine these architectures, resulting in the CNNLSTM model. It is also referred to as the encoder-decoder model. The neural network-based image captioning methods work in

a straightforward end-to-end fashion. These methods are very similar to neural machine translation based on the encoder-decoder framework. Global image features are extracted from CNN hidden activations and fed into an LSTM to generate a word sequence in this network. The CNN-LSTM architecture is constructed by combining CNN layers for feature extraction on input data with LSTMs for sequence prediction. This model is intended for sequence prediction problems with spatial inputs, such as images or videos. They are widely used in activities such as activity recognition, image description, video description, and many others. CNN-LSTMs are typically used when their inputs have spatial structure, such as the 2D structure of pixels in an image or the 1D structure of words in a sentence, paragraph, or document, as well as temporal structure, such as the order of images in a video or words in text, or when output with temporal structure is required, such as words in a textual description. We are introducing more and more controlling knobs as we use LSTM over RNN, which controls the flow and mixing of inputs based on trained Weights. As a result, more control over the outputs is possible. As a result, LSTM provides the most control and thus better results. CNN-RNN Model: Using CNN for Object Detection: CNN yields optimistic results for object detection and is best suited for image captioning. RNN-LSTM will be used to generate meaningful captions from image and object detection features. The object detection will be the input, and the output will be the caption for the specific image. Image captioning has improved significantly in recent years. The neural image caption generator provides a useful framework for learning how to map different images to human-level image captions. Using tensorflow and algorithms, neural networks can handle all of the issues by generating appropriate, expressive, and highly fluent captions. The efficiency of content-based image retrieval can be improved by text description of the images, as well as the expanding application scope of visual understanding in science, security, defence, and other fields, which has a wide application prospect. This Image Captioning deep learning model is extremely useful for inspecting large amounts of unstructured and unlabelled data in order to detect patterns in those images for guiding self-driving cars and developing software to assist blind people.



3.6 Result

We greedily choose the word with the highest probability to get the next word prediction as the model generates a 1015-long vector with a probability distribution across all the words in the vocabulary. Optimal Search is the name of this approach.In prediction search, the top k predictions are taken, fed back into the model, and then sorted based on the probabilities the model returned. Therefore, the top k predictions will always be included in the list. We then select the prediction with the highest probability and proceed through the list until we reach ”last” or the maximum caption length.Let’s try out our model now and see what captions it comes up with for various photographs. We’ll also examine the various captions produced by the optimal search and the prediction search with various k values.



4 Conclusion

We reviewed deep learning-based image captioning methods in this paper. We provided a taxonomy of image captioning techniques, displayed generic block diagrams of the major groups, and highlighted their advantages and disadvantages. We discussed various evaluation metrics

and datasets, as well as their advantages and disadvantages. A summary of the experimental results is also provided. We provided a brief overview of potential research directions in this area. Although deep learning-based image captioning methods have made significant advances in recent years, a robust image captioning method capable of producing high-quality captions for nearly all images has yet to be developed. With the introduction of novel deep learning network architectures, automatic image captioning will continue to be an active research area for some time. The text file also contains the captions for the almost 8000 photographs that make up the Flickr 8k dataset that we used. Although deep learning-based image captioning techniques have made significant strides in recent years, a reliable technique that can provide captions of a high caliber for almost all photos has not yet been developed. Automatic picture captioning will continue to be a popular study topic for some time to come with the introduction of novel deep learning network architectures. With more people using social media every day and the majority of them posting images, the potential for image captioning is very broad in the future. Therefore, they will benefit more from this project.

5 Future Scope

Upcoming tasks The exponential rise of photos on social media and the internet in recent years has made image captioning a significant issue. This study outlines the many methods and approaches employed in the research while also discussing the various picture retrieval research conducted in the past. Future study in this area has a huge potential because feature extraction and similarity calculation in images are difficult tasks. IMAGE RETRIEVAL USING IMAGE CAPTIONING 54 Histogram, color, tags, and other features are used in current image retrieval systems to calculate similarity. Results cannot be totally accurate because these approaches are independent of the image's context.Therefore, a thorough study of picture retrieval using the image context, such as image captioning, will help to resolve this issue in the future. By including new picture captioning datasets into the project's training, it will be possible to improve the identification of classes with lesser precision in the future. In order to see if the picture retrieval outcomes improve, this methodology can also be integrated with earlier image retrieval techniques like the histogram, shapes, etc.

6 References

1. William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation. arXiv preprint arXiv:1801.07736, 47, 2018.

2. Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.

3. Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. Workshop on Neural Information Processing Systems (NIPS), 2014.

4. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning, 2048- 2057, 2015.

5. Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, pages 1057–1063, 2000.

6. Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.

7. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model.