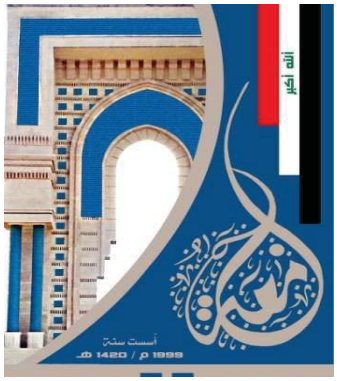

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة ديالى
كلية العلوم
قسم علوم الحاسبات

مولد تعليق الصور باستخدام نماذج التعلم العميق

رسالة مقدمة
الى كلية العلوم في جامعة ديالى وهي جزء من متطلبات نيل
شهادة الماجستير في علوم الحاسبات

تقدم بها الطالب

ياسر حميد زيدان

بإشراف

أ.م.د جمانة وليد صالح

١٤٤٤ هـ

٢٠٢٣ م

Chapter One

Introduction

Chapter One

General Introduction

1.1 Introduction

Automatically producing captions or descriptions for photographs is a challenging subject that involves a visual combination & language input. In other words, it requires both comprehensive visual understanding and advanced natural language creation. The communities of computer vision and natural language processing have embraced it as an intriguing problem [1]. image captioning is the process of automatically generating a description of an image in natural language. The study of image captioning, a bridge between computer vision and natural language processing research, requires not only a thorough comprehension of a picture semantic contents but also the capacity to communicate the information in human-like sentences. Recognizing the existence, qualities, and connections of items in an image is tough enough. The complexity of structuring a phrase to convey such information adds to the task difficulty [2].

Researchers have been attempting to use a plausible sentence in English or another language to describe the content of an image in recent years, which has gained popularity in the field of computer vision. A model for automatic image caption generation frequently transforms the objects, semantic attributes, object relationships, and predicted actions contained inside it into a representation vector in order to describe an image. On top of that, the caption may be produced word by word utilizing a word generator. As a consequence of recent developments in deep learning techniques, a great number of approaches to that challenge have been published employing that paradigm [3].

The research community has made significant strides in model design over the last few years: The methods have been improved with attentive approaches and reinforcement learning in recurrent from the initial proposals based on deep learning and using to the developments of Transformers and self-attention, as well as single-stream like (Bidirectional Encoder Representations from Transformers) techniques. Visual characteristics that are global in scope are fed to Neural Networks (NNs). Simultaneously, researchers in the disciplines of computer vision and natural language processing have developed assessment processes and criteria for comparing outcomes to human-generated ground facts. Despite years of research and advancements, The issue of image captioning has not yet been fully resolved [4].

A number of studies have been done on automatic images captioning. It can be broken down into the following three categories: creating new image captions, retrieval-based image captioning, and template-based image captioning. A fixed template is filled in by searching for objects, attributes, and actions before using template-based picture captioning. In the training dataset, retrieval-based techniques first identify visually comparable photos and their captions. The caption for the image is then selected from a group of captions for related pictures. These methods can make captions that are correct in terms of grammar, but not in terms of what they mean. On the other hand, the new ways of making image captions look at the image's visual content and use a language model to make image captions based on what it sees. For an image, novel caption generation can produce new captions that are more logical than the ones that came before, in contrast to the previous two categories. Most works in this area make use of machine learning and deep learning. The encoder-decoder framework for image captioning is a well-liked framework in this area. Deep Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were used

as the encoder and decoder, respectively, in recent studies. But it's still hard to find the best CNN and RNN models for image captioning [5].

Images are a rich and expressive medium, but they lack the explicit textual information that is readily available in written language. By generating captions or descriptions for images, we can provide additional context, comprehension, and accessibility to visually impaired individuals or those who prefer textual information. The problem in image captioning refers to the difference between human-level understanding and the current performance of automated systems in generating accurate and semantically meaningful descriptions for images. Generating accurate and meaningful captions requires a deeper understanding of the visual scene, relationships between objects, and the ability to generate fluent and contextually relevant descriptions.

This proposed system addresses the challenge of connecting visual information with natural language understanding. By generating textual descriptions for images, it helps bridge the gap between visual perception and linguistic comprehension, contributing to a deeper understanding of the relationship between images and language.

1.2 Related Works

In this section, the study examines certain deep learning strategies and techniques that make use of computer vision and Natural Language Processing (NLP) methods. Some of these are briefly explained below:

-Qiu et al. (2020) [6] developed an approach for resolving the issues of onboard image priority for data downlinks and local images search for planetary images data server by integrating the responsibilities of images captioning and text similarity evaluations. This combination of tasks is intended to solve the problem of images captioning and text similarities

evaluation. That proposed approach includes several stages, firstly, feature extraction, then visual attention, finally, text decoder. In order to evaluate that approach, the Martian Image caption Dataset (MICD) was utilized. From experiments, the obtained accuracy was about 80%.

-Wang et al. (2020) [7] presented a method for image captioning that includes two new components: graphs-based visuals relationship modeling and a context-aware attentions mechanism. The model uses a convolutional neural network (CNN) to extract images feature, a generative neural network (GNN) to learns the implicit visual relationships among the visual objects or regions in an images, a visual context-aware attention model to select relevant relationships from among 160 representations, and LSTM based languages model to generate sentences. This strategy was examined through the lens of MS COCO and Flickr30K.

-Yan et al. (2020) [8] recommended the use of a method known as hierarchical attention, which would increase performance by accounting for both the features of the detected items and the global image features. The CNN encoder and the object detector in this type of setup are in charge of obtaining global and local information, respectively. These two categories of features are input into the LSTM models through the application of global and local attention strategies. Concatenating and decoding the outputs of two different LSTM models results in the formation of words. In order to evaluate the model, the MS COCO dataset was utilized.

-Zhang et al. (2021) [9] introduced a unique visual connection attention model based on parallel attention and learning spatial restrictions for the first time. In that model, the image encoder is a Faster R-CNN, and the language decoder is a two-layer LSTM. Both attention models can be found in the LSTM-1 (which acts as a coarse decoder) and the LSTM-2 (which acts as a

fine decoder) (fine decoder). In order to test this method, the MS COCO dataset was utilized.

-Li et al. (2021) [10] introduced a novel multilevel similarity guided semantic matching technique for images captioning that can learn the latent semantic connection between images and create captions through merging local and global semantic similarities. That model comprises of visual semantic unit extraction and encoding using Faster R-CNN, as well as languages decoding using LSTM and an attentions mechanism. MS COCO dataset was utilized to evaluate the model.

-Wang and Gu (2022) [11] used the global and local visual cooperation attention strategy to examine the natural relationships between global and local picture components. A new network specifically was developed using fusion modules and a visual interaction encoder. The first module focused on the implicit recording of visual relationships between local and global image components in order to provide a more accurate rich depiction. While the second module worked on combining the previously acquired features to obtain extra better information on various-level relation property. Another module (LSTM) was included as well to control word formation. The extensive experiment results demonstrated the superiority of the suggested strategy used on the MS COCO dataset.

-Verma et al. (2023) [12] presented a system using the CNN-LSTM architecture to create an image caption, or description of an image so that the current word serves as an input for the prediction of the next word. CNN layers (VGG16 Hybrid Places 1365) assist in extracting data from the input, and LSTM extracts necessary details throughout the processing of input to create sentences that describe the content of the image. Two datasets (Flike8K and MSCOCO) were utilized to evaluate this proposed system.

Table 1.1: some of the approaches to image caption using deep learning models

Author/(s), Ref, Year.	Deep learning models	Datasets	BLEU				METEOR
			B1	B2	B3	B4	
Qiu et al. (2020) [6]	CNN, RNN	MICD	0.547	0.482	0.431	0.388	0.309
Wang et al. (2020) [7]	CNN, GNN , LSTM	MS COCO	0.759	0.603	0.465	0.358	0.278
		Flickr30K	0.698	0.517	0.378	0.277	0.215
Yan et al. (2020) [8]	CNN , LSTM	MSCOCO	0.73036	0.53688	0.39069	0.285	0.25324
Zhang et al. (2021) [9]	Faster R-CNN, LSTM	MSCOCO	0.788	0.614	0.472	0.363	0.279
Li et al. (2021) [10]	Faster R-CNN, LSTM	MSCOCO	0.812	-	-	0.390	0.285
Wang and Gu (2022) [11]	LGVIA, LSTM	MSCOCO	0.814	0.651	0.505	0.389	0.285
Verma et al. (2023) [12]	CNN, LSTM	MSCOC	0.7350	0.542	0.423	0.334	0.4768
		Flickr8k	0.6666	0.4704	0.389	0.287	0.506

A set of datasets has been detailed in Table 2. These datasets may be viewed as a continual endeavor on the part of researchers to offer the massive quantities of diverse data required for the most advanced deep-learning neural networks.

Table 1.2: Datasets of image captioning

Dataset	Ref.	No. image	Description
Flickr 8K	Hodosh et al. [13]	8,000 images	Humans and animals are the most common subjects in that dataset.
Flickr 30K	Peter Young et al.[14]	31,783 images	Mostly about people going about their daily lives and events.
MSCOCO	Xinlei Chen et al.[15]	123,287 images	To produce that project, images of complicated everyday scenarios with ordinary things in their natural contexts were gathered.
Sydney	Bo Qu et al.[16] , Fan Zhang et al. [17]	613 images	The seven classes based on the Sydney Dataset include runway, industrial, residential, airport, grassland, rivers, ocean, and oceanfront.
UCM	Bo Qu et al.[16] , Yi Yang et al. [18]	2,100 images	A baseball diamond, a beach, a building, chaparral, dense residential, a harbour, a junction, a medium residential area, a mobile home park, an overpass, a parking lot, a river, a runway, and a tennis court are just a few of the 21 categories that are available.
RSICD	Xiaoqiang Lu et al.[19]	10,921 remote sensing images	The collection contains photos obtained from Google Earth and scaled to 224 224 pixels at a variety of resolutions.
MICD	Dicong Qiu et al. [6]	12,500 images	Capture Martian geologic characteristics, particularly landscape features, in images that contain several item or feature categories.

1.3 Problem Statement

The problem proposes a captioning job that calls for a computer vision system to identify and characterize in natural language significant regions in images. When the descriptions are just one word long, the image captioning task generalizes object detection. Find the appropriate

semantic label for each image in a collection of photographs using knowledge of the content.

Images captioning is still a very challenge task since it mixes computers vision and natural languages processing. In other words, it requires both comprehensive visual understanding and advanced natural language creation. In addition to having a thorough comprehension of the semantic content of the image, the information in an image caption must be expressed in a manner that is human-like. Finding things in an image and determining their presence, traits, and relationships is challenging enough. The complexity of structuring a phrase to convey such information adds to the task's difficulty.

Natural language processing (NLP) created an RNN-based model that is used to predict the sequences of words, called the caption, from the feature vectors obtained from the pre-trained CNN network.

1.4 Aim of the thesis

This thesis major goal is to produce captions for photographs that are of the highest quality and contain accurate and useful item information.

To achieve this aim, several objectives will be done:

- Providing deep networks with more tools, such as Inception, with attention to producing insightful and excellent image captions.
- Including local, past, and future context for creating semantically rich image captions.
- Apply an end-to-end deep learning-based image captioning scheme using pre-trained CNN (EfficientNet-B7) and Long-Short Term Memory (LSTM) with visual attention. The model of LSTM with visual attention process the vector of an image feature in order to fine-

grain and further abstract visual depiction. The language-LSTM is capable of keeping salient objects and predicting the following word on the track of context words and objects.

- Integrate the accommodative attention model for extracting vision features, mining language structure, and generating image descriptions with more details.

1.5 Thesis Outlines

The study is divided into five chapters, the following is a brief summary of what each chapter covers:

Chapter Two: Explains the theoretical underpinnings of the commonly used methods for creating image captions, along with the benefits and drawbacks of each type of method.

Chapter Three: Declares the methodology of the proposed systems.

Chapter Four: Gives a description of the experiments that are carried out in order to assess the suggested systems and verify the hypothesis of this work, as well as the data gathered from these tests.

Chapter Five: Provides some recommendations for future works and conclusions.

الخلاصة

التسمية التوضيحية للصورة هي عملية استخدام نظام الفهم البصري مع نموذج للغة ، يمكننا من خلاله بناء جمل ذات معنى ودقيقة نحويًا للصورة. الهدف هو تدريب نموذج التعلم العميق لمعرفة التوافق بين الصورة ووصفها النصي. هذه مهمة صعبة بسبب التعقيد المتأصل والذاتية للغة ، وكذلك التباين المرئي للصور. لذلك ، يجب أن يتضمن أي مخطط تسمية توضيحية للصورة القدرة على تحليل وفهم الدلالات المرئية للصورة من خلال إدراك المكونات البارزة وتفاعلها / ارتباطها. يتم استخدام كل من رؤية الكمبيوتر ومعالجة اللغة الطبيعية في المهمة الصعبة المتمثلة في تسمية الصور.

تم اقتراح نظامين باستخدام نموذج Encoder-Decoder ، النظام الأول يستخدم EfficientNet-B7 الذي تم تدريبه مسبقًا والثاني يستخدم Inception V3 مُدرَّبًا مسبقًا باعتباره جهاز تشفير لاستخراج الميزات ، ويستخدم النظامان LSTM مع آلية الانتباه كوحدة فك ترميز لإنشاء تسميات توضيحية كلمة بكلمة مع التركيز على الأجزاء الأكثر صلة بالصورة. تسمح آلية الانتباه للنموذج بالحضور إلى أجزاء مختلفة من الصورة في أوقات مختلفة أثناء عملية إنشاء التسمية التوضيحية ، مما يؤدي إلى تسميات توضيحية وصفية وأكثر دقة.

يتم تدريب الأنظمة المقترحة على مجموعة بيانات MSCOCO (كائنات Microsoft المشتركة في السياق) باستخدام مقاييس Bleu (B1 و B2 و B3 و B4) ودرجة Meteor (مع تدريب بنسبة 80% واختبار 20%). حقق النظام باستخدام EfficentNet-B7 النتائج (B1 = 0.888، B2 = 0.875، B3 = 0.857، B4 = 0.666 ، و Meteor = 0.698) ، وحقق النظام مع Inception V3 النتائج (B1 = 0.78، B2 = 0.65، B3 = 0.51، B4 = 0.42 ، و Meteor = 0.543). EfficientNet-B7 هي بنية شبكة عصبية تلافيفية أكثر تقدمًا وقوة مقارنةً بـ Inception V3 ، وتشير النتائج إلى أنها أكثر فاعلية في تسمية الصور على مجموعة بيانات MSCOCO في هذه التجربة بالذات.