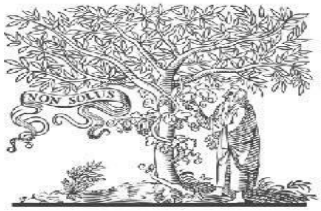
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 10th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

10.48047/IJIEMR/V12/ISSUE 04/97

Title IMAGE CAPTIONING USING REAL AND SYNTHETIC IMAGES

Volume 12, ISSUE 04, Pages: 784-790

Paper Authors

Mrs. M. Rajya Lakshmi, A.Supriya, B.Vaishnavi, CH.Sunandini, K.Keerthana

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

IMAGE CAPTIONING USING REAL AND SYNTHETIC IMAGES

Mrs. M. Rajya Lakshmi, Associate Professor, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

A.Supriya, **B.Vaishnavi**, **CH.Sunandini**, **K.Keerthana**
UG Students, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
aisanisupriya17@gmail.com, badakerevaishnavi@gmail.com,
nandusrinivas16@gmail.com,glorykeerthana666@gmail.com

Abstract

Computer vision and natural language processing have placed significant emphasis on producing textual descriptions for images over an extended period of time. Numerous methods based on deep learning have been established to achieve this objective, among which Image Captioning is the most conspicuous. The key objective of this project is to generate descriptions for images submitted by the user. This is achieved through a Python-based implementation of caption generation using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models. These models are trained and tested using both human-annotated images and synthetic data generated by a Generative Adversarial Network (GAN)-based text to image generator. The task of collecting a substantial number of human-generated images along with their corresponding descriptive captions can be both costly and time-consuming, but our proposed method has overcome this limitation by using synthetic data. The evaluation of the models involved commonly utilized qualitative and quantitative analyses, which demonstrated a substantial improvement in the quality of the descriptions generated for real images when both real and synthetic data were used during the training phase.

Keywords: GAN, LSTM, CNN, Corpus Text, Lemmatization, Deep Learning.

1.Introduction

The task of creating a written explanation of an image is referred to as image captioning. It involves understanding the content of an image and expressing that understanding in natural language. Generating textual descriptions for images is a complex task that necessitates the use of both computer vision and natural language processing techniques.

Real images are photographs or digital images that have been taken in the real world, while synthetic images are computer-generated images that are designed to look like real images. Image captioning can be performed on both real and synthetic images.

Image captioning using real images involves training a model on a large dataset of real images and their

corresponding captions. This dataset can be obtained from various sources, such as Flickr or COCO (Common Objects in Context). The model learns to associate features extracted from the image with words and phrases in the caption. The caption will be generated by model for a new image by extracting features from the image and using them to predict the corresponding caption.

Image captioning using synthetic images involves generating images using computer graphics software and training a model on these synthetic images and their corresponding captions. Synthetic images can be generated faster and more easily than real images, which can be useful for applications where a large number of images are required.

Figure 1: sample image

If we were asked to describe a scene involving a puppy on a blue towel or a brown dog playing with a green ball, we could easily do so as humans. Automatic generation of a textual description for an image with an artificial system is inherent process in Image captioning.

Assessing the model's accuracy is critical, given that images can be described in various ways. The BLEU score is a measurement employed to evaluate how closely the generated sentence matches the reference sentence. It is commonly used for sequence-to-sequence problems, such as summarization, language translation, and captioning, with a perfect match having a score of 1.0 and a perfect mismatch having a score of 0.0.

1.1 Image synthesis

Image synthesis refers to the creation of artificial images with specific content, similar to solving the inverse classification problem of generating an image with specific label-associated visual contents. One way to generate synthetic images is through generative adversarial networks (GANs). GANs comprise two CNNs that are trained in a competitive manner. The generator CNN generates synthetic images to deceive the discriminator CNN, which distinguishes between real and synthetic images. GANs are widely used in various domains to generate synthetic images based on captions.

Generator:

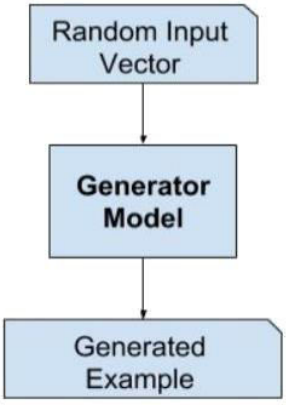


Figure 2: image generator

The generator in a deep learning model plays a crucial role as it is responsible for producing synthetic examples that meet certain criteria. During the training process, it is essential to focus on improving the generator's performance to ensure that it can generate high-quality synthetic examples. The generator's primary objective is to produce synthetic examples that match a specific input. For instance, If the training data for the generator consists of cat images, it ought to be competent in producing a convincing portrayal of a cat.

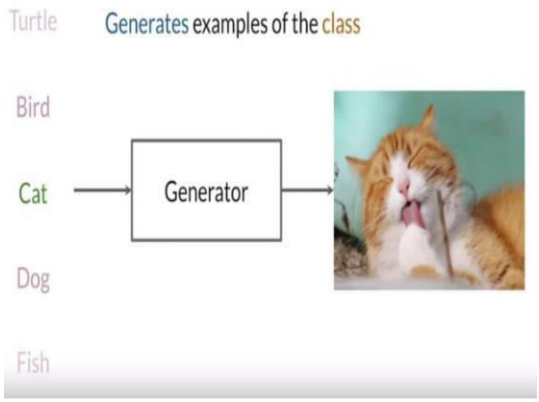


Figure 3: sample image generation

2. Literature Survey

Through the studies on the use of real and synthetic images for generating captions found to be an active area of research. Some of the key research works are addressed here.

In a 2016 paper, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Xu et al. proposed an image captioning model that uses both synthetic and real images. The authors used synthetic images generated by rendering 3D models as well as real images from the COCO dataset. The authors demonstrated that their model outperformed previous approaches on several evaluation metrics.

In a 2017 paper, "Towards Diverse and Natural Image Descriptions via a Conditional GAN," Chen and Zitnick proposed a conditional generative adversarial network (GAN) for image captioning that used both synthetic and real images. The authors used synthetic images generated by a GAN to augment their real image dataset and demonstrated that their approach improved the diversity and naturalness of the generated captions.

In a 2018 paper, "Learning to Learn Image Captioning with Uncertainty," Li et al. proposed an image captioning model that used synthetic images to improve performance on real images. The authors used synthetic images generated by a GAN to train their model with uncertainty sampling, a technique that selects samples with the highest prediction uncertainty for training. The authors demonstrated that their approach improved the performance of their model on real images compared to a model trained only on real images.

In a 2019 paper, "Synthetically Supervised Feature Learning for Scene Understanding," Wang et al. proposed a method for generating synthetic images to improve feature learning for image captioning. The authors used a GAN to generate synthetic images and trained their model on both synthetic and real images. The authors demonstrated that

their approach improved the performance of their model on several evaluation metrics.

In a 2020 paper, "Synthetically Supervised Image Captioning," Münch et al. proposed a method for generating synthetic images to improve performance on image captioning. The authors used a GAN to generate synthetic images and trained their model using a synthetically supervised loss function that combined the predictions of a real-image-based model and a synthetic-image-based model. The authors exhibited that their method enhanced the efficiency of their model for both real and artificial images.

In general, these investigations indicate that utilizing synthetic images in image captioning can enhance the quality and variety of the produced descriptions without any plagiarism. However, there is still much work to be done in this area, particularly in developing more efficient and cost-effective methods for generating synthetic images.

3. Problem Statement

1.How can captions be generated from images?

2.Can the performance of image captioning be improved using synthetic images?

3.What is the process of generating synthetic images?

Currently, image captioning is a widely researched area; however, existing caption generators are only trained on labeled authentic images. To enhance the performance of caption generators, it is imperative to develop caption generators that can work with both real and synthetic images. This study aims to broaden the training of caption generators by incorporating both real and synthetic images. The study employed Generative Adversarial Networks (GANs) to produce synthetic images from input text, creating a dataset of synthetic images with corresponding factual captions. These synthetic images were then utilized in combination with genuine images to train and test an image captioning model. Ultimately, the study demonstrated that incorporating synthetic images can significantly enhance the quality of the produced captions.

4. Methodology

BigGAN: The BigGAN is a kind of generative adversarial network that has been created to facilitate the generation of high-resolution, high-fidelity images at scale. It is a specialized model for class-conditional image generation, which enables the production of larger and superior quality images, such as those of 256×256 and 512×512 resolution.

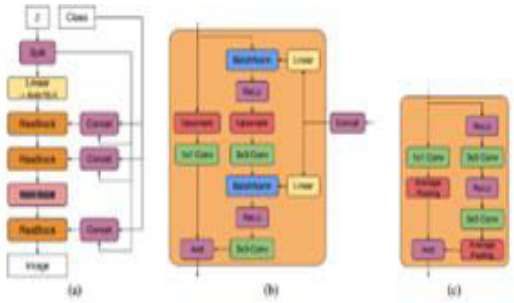


Figure 4: (a) A typical architectural layout for BigGAN's G (b) A ResBlock up in BigGAN's G (c) A ResBlock down in BigGAN's G

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

VGG16 model:

The VGG16 is a Convolutional Neural Network (CNN) that is widely regarded as one of the most effective computer vision models available. With its ability to accurately classify 1000 photos from 1000 different categories at an impressive rate of 92.7%, VGG16 is a reliable method for object identification and classification. It is a favoured option for image classification and can be conveniently executed using transfer learning techniques.

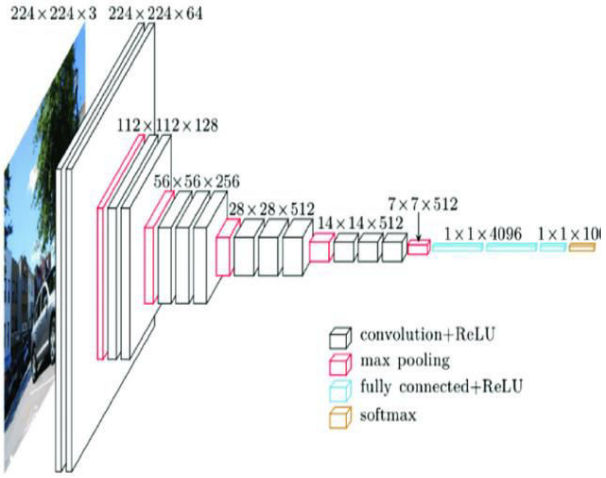


Figure 5: Architecture of VGG Model

LSTM model:

To describe long short-term memory networks, which are used in deep learning, the term "LSTM" is used. LSTMs are a type of recurrent neural network (RNN) that are effective at learning long-term relationships, particularly in sequence prediction tasks.

LSTMs excel in managing long-term dependencies, mainly because they can maintain information over extended periods of time. In an LSTM, a set of "gates" is employed to manage the input, storage, and output of data in a sequence. The forget gate, input gate, and output gate are the three gates that make up a standard LSTM. These gates are individual neural networks that function as filters.

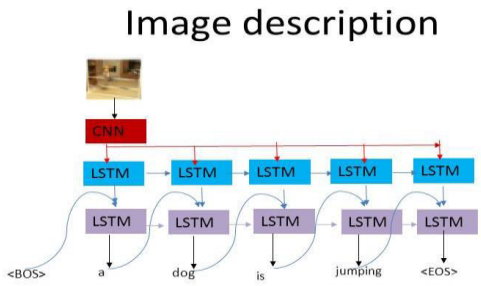


Figure 6: Illustration of image description generator.

5. Implementation

5.1 Preprocessing

To extract useful insights from real-world data, we need to perform some basic cleaning since the data is often dirty. This involves procedures such as lower-casing all words to ensure that similar words are treated as the same, removing special tokens like '%', '$', '#', and so on, as well as eliminating words that contain numbers such as 'hey199'.

In caption preprocessing, the following steps were taken:

- All words with only one letter were removed.
- All special characters were removed
- Tags like 'startseq' and 'endseq' were added to indicate the start and end of a caption for easier processing.

5.2 Feature Extraction

Feature Extraction includes the following steps:

1. Loading the image from a file.
2. Converting the image pixels into a numpy array.
3. Reshaping the data to suit the model's requirements.
4. Preprocessing the image for VGG.
5. Extracting features from the image.
6. Storing the extracted features.

5.3 Dataset

The FLICKR 8K dataset was obtained from Kaggle and is accessible to the public. The dataset comprises 8,000 images, each with five captions. These images were selected from six Flickr groups and do not typically feature recognizable individuals or locations, but rather depict a range of scenarios and environments. The dataset comprises of 8,000 distinctive images, with each image being linked to five exclusive sentences.

.

5.4 Metrics Calculated

To assess the performance of different machine learning algorithms, metrics are commonly employed. In this particular model, the BLEU score is utilized as the evaluation metric.

Bleu Score:

5.3.1 Precision Score

To illustrate the procedure, let us consider an NLP model that generates a predicted sentence for a single target sentence.

Target Sentence: The guard arrived late because it was raining.

Predicted Sentence: The guard arrived late because of the rain.

Although the approach for multiple target sentences is similar, we will only focus on one for simplicity. The initial step is to calculate the precision scores for n-grams ranging from 1 to 4.

Precision 1-gram

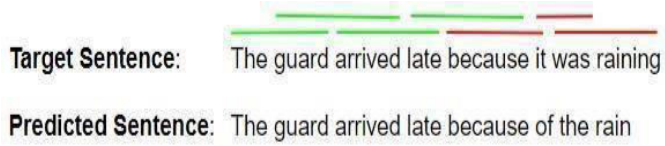
We utilize the Clipped Precision approach.

Precision 1-gram = (Number of correct predicted 1-grams)/(Number of total predicted 1-grams)

So, Precision 1-gram (p_1) = 5 / 8

Precision 2-gram

Precision 2-gram = Number of correct predicted 2-grams / Number of total predicted 2-grams.



So, Precision 2-gram (p_2) = 4 / 7

$$Bleu\,(N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

6. Results

The training of the model involved both real and synthetic images, which was then contrasted against a model that was exclusively trained using real images.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Figure 7: caption generated using real images for training.

Figure 8: caption generated using real and synthetic images for training.

7. Conclusion

The obtained results indicate that training the model with both synthetic and real images leads to an enhancement in the quality of generated captions.

The corresponding BLEU scores are:

Model	BLUE-1	BLUE-2
With real images	0.523	0.291
With real and synthetic images	0.544	0.319

8. Future Work

The current study utilized text-only generated synthetic images. However, generating synthetic images from real images could be a potential avenue for future research. Additionally, exploring the use of synthetic captions to further improve image captioning could also be an interesting area for further investigation.

9.References

[1]. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 2758–2766.

[2]. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in Proc. Eur. Co nf. Comput. Vis., 2010, pp. 15–29.

[3]. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3128–3137.

[4]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[5]. B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Laze bnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 2641–2649.