From Pixels to Words: Tracing the Journey of Automated Image Captioning

by Debashree Chakraborty

Submission date: 10-Apr-2024 04:23PM (UTC+0530)
Submission ID: 2345462517
File name: Debashree___Aditi_8thSem.pdf (3.8M)
Word count: 2990
Character count: 18430

From Pixels to Words: Tracing the Journey of Automated Image Captioning

Debashree Chakraborty
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 20051139@kiit.ac.in

Aditi Singh
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: 20051629@kiit.ac.in

Santos Kumar Baliarsingh
School of Computer Engineering
KIIT Deemed to be University,
Bhubaneswar, India 751024
Email: santos.baliarsinghfcs@kiit.ac.in

***Abstract*—Image caption generation is a critical research area that combines computer vision and natural language processing, with wide-ranging implications such as assisting visually impaired individuals, improving autonomous vehicle capabilities, and refining image search algorithms. The main goal of image captioning is to identify objects in an image and explain the relationships between them using text. In the context of image captioning, this research compares three popular encoding architectures: ResNet50, VGG16 and InceptionV3. To analyse the input image in order to obtain significant features that are then converted for vector representation, an Encoder component of the architecture is used. These encoded features are used by the decoder module to create a coherent text description using the Long Short Term Memory LSTM model. By conducting thorough experiments and evaluations, this study aims to uncover the performance differences among the mentioned encoder architectures and determine the best model for image captioning tasks.**

***keywords*—Image captioning, Encoder-decoder architecture, ResNet50, VGG16, InceptionV3, and LSTM Comparative analysis.**

I. Introduction

A. Background

Every day, we encounter a plethora of images sourced from various platforms, ranging from the vast expanse of the internet to news articles, data charts, and advertisements. While many images may convey implicit meaning, detailed captions play a crucial role in enhancing communication and accessibility, serving as a bridge between visual content and interpretation.

The need for automated captioning arises from the fact that machines require structured data to interpret images effectively, unlike humans who can infer meaning from context. Captions provide this structured data by offering textual representations of image content, thereby improving the efficiency of image search and indexing processes. Additionally, comprehensive captions provide deeper understanding compared to simplistic object recognition approaches, aligning more closely with human interpretation.

Achieving natural language generation by machines poses significant challenges, requiring capabilities in complex natural language processing and understanding. Despite these challenges, the applications of captions span diverse domains such as biomedicine, commerce, web search, and military operations. In biomedicine, captions aid in interpreting medical images for diagnosis and treatment planning, while in commerce, they enhance product descriptions and marketing materials, fostering consumer engagement.

Captions also play a crucial role in web search algorithms, enabling more accurate indexing and retrieval of visual content. Moreover, in military contexts, captions contribute to situational awareness and intelligence analysis. The prevalence of social media platforms further underscores the importance of automated caption generation, as platforms like Instagram and Facebook automatically generate captions for uploaded images, enhancing accessibility and user engagement by providing textual context alongside visual content.

B. Motivation

The research is motivated by the need to provide image explanations, which is a critical aspect of computer vision and natural language processing. Although machines have made progress in generating human-like descriptions, accurately capturing the complex relationships between objects in images and expressing them fluently in languages like English remains a challenge. Conventional methods that rely on predefined templates lack flexibility and often fail to produce linguistically rich descriptions.

In order to address these limitations, researchers have turned to neural networks, harnessing their ability to learn intricate patterns and connections. Modern image description models leverage neural networks to create coherent and contextually relevant descriptions without being constrained by fixed templates, resulting in significant improvements in the quality and adaptability of image representation systems. Through continuous refinement of neural models, researchers strive to enhance their capacity to generate precise and informative descriptions, ultimately aiming to develop machines capable of understanding and describing visual content with human-like proficiency. This endeavor opens up new possibilities for AI-driven applications across a wide range of domains.

C. Contributions

Our study thoroughly compares and contrasts three well-known encoder architectures: ResNet50, VGG16, and InceptionV3, which significantly advances the field of image captioning. Through a meticulous evaluation of these architectures, we offer valuable insights into their performance in generating coherent textual descriptions from images. Our study not only examines traditional metrics like caption accuracy and fluency but also delves into aspects such as computational efficiency and scalability, providing a comprehensive view of their effectiveness in real-world scenarios. Through rigorous experimentation and evaluation, our goal is to pinpoint the optimal model for image captioning tasks, which can serve as a valuable resource for both researchers and developers.

Furthermore, our research sheds light on the intricate relationship between the encoder and decoder components of the image captioning framework. By emphasizing the practical implications of our findings, we bridge the gap between theoretical research and practical applications, empowering a wide range of stakeholders—from individuals with visual impairments to developers of autonomous systems—with transformative image captioning technology. In summary, our project pushes the boundaries of image captioning technology by offering empirical evidence, insights, and recommendations to enhance its effectiveness and applicability across diverse domains.

II. LITERATURE REVIEW

The literature review synthesizes findings from eleven research papers, encompassing various aspects of image captioning, including encoder architectures, attention mechanisms, dataset considerations, and evaluation methodologies. These papers collectively contribute to advancing the current understanding of image captioning by elucidating the strengths and limitations of different approaches, thereby identifying avenues for further research and improvement.

Several overarching themes and patterns emerge from the literature. Firstly, the importance of encoder architectures, such as VGG16, ResNet, and InceptionV3, is underscored in extracting meaningful features from images. Studies like [1], [2], [3], and [4] delve into the comparative analysis of these architectures, highlighting the significance of selecting appropriate models for effective caption generation. Furthermore, attention mechanisms play a crucial role in enhancing the focus on salient image regions, as discussed in papers like [2], [5], [6], and [7], thereby improving the quality of generated captions.

Dataset considerations also emerge as a critical aspect, with studies like [8], [9], and [7] emphasizing the importance of comprehensive datasets tailored to specific domains, such as remote sensing imagery or incorporating external knowledge. Moreover, evaluation metrics like BLEU score and qualitative analysis are highlighted across multiple papers ([1], [2], [3], [6]), demonstrating the need for trustworthy assessment methods in order to accurately determine the efficacy of image captioning models.

However, the literature also reveals several gaps and inconsistencies. For instance, while attention mechanisms are shown to improve caption accuracy in some studies ([2], [5], [6]), challenges in interpreting attention's role and its correlation with caption accuracy remain unclear ([7]). Additionally, limitations in existing datasets ([6], [9]), occasional inaccuracies in generated captions ([10], [3], [4]), and scalability issues ([5],[6]) underscore the need for further exploration and refinement in image captioning research.

These insights can inform the direction and objectives of future research projects. For instance, addressing the interpretability of attention mechanisms and their impact on caption quality could lead to more robust and transparent image captioning systems. Furthermore, efforts to construct comprehensive datasets tailored to specific domains, incorporating external knowledge, and improving evaluation methodologies can enhance the reliability and generalizability of image captioning models. Additionally, exploring advanced techniques such as reinforcement learning, unsupervised learning, and fusion of textual and visual information may further advance the state-of-the-art in image captioning technology, addressing existing limitations and fostering innovation in the field.

III. PROPOSED WORK

Image captioning methods encompass template-based, retrieval-based, and novel caption generation approaches, while deep learning-based methods primarily rely on visual and multimodal spaces. Encoder-decoder architectures use CNNs for image feature extraction and LSTMs for sequence generation, while compositional architectures employ attention mechanisms or reinforcement learning. Deep learning techniques have shown promising results, leveraging CNNs to capture image features and LSTMs to generate accurate captions by understanding long-range dependencies in sequences. Overall, these methods have revolutionized image captioning by automating the process and improving caption quality. We employ CNN and LSTM for picture categorization. So, to build our photo caption generating model, we'll mix these designs. It is sometimes called the CNN-RNN model.

- CNN extracts features from images.
- LSTM will use CNN data to create an image description.

A. *Convolutional Neural Network*

Convolutional neural networks, often known as CNNs, are deep learning algorithms that can recognize distinct objects and characteristics in an input picture and determine their relative significance using learnable weights and biases. When it comes to pre-processing, ConvNet requires a lot less than other classification methods. One type of unique deep neural network that can handle data in input forms like 2D matrices is the convolutional neural network. CNNs are highly helpful when working with photographs, because images may be represented as a 2D matrix with ease. To extract significant details from a photograph, scan it from top to bottom and left to right. Finally, integrate the features to classify the image. Images that have been scaled, rotated, altered, and transformed can be processed in perspective.

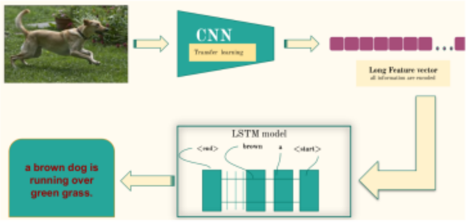


Figure 1. Proposed Architecture of Convolutional Neural Network.

B. *Long Short- term Memory*

LSTM, brief for Long Short-Term Memory, is a sort of repetitive neural arrange (RNN) well-suited for grouping expectation assignments, such as foreseeing the following word in a sentence. Overcoming the restrictions of conventional RNNs' short-term memory, LSTM has demonstrated more compelling by keeping up significant data all through input preparing and utilizing disregard doors to dispose of unessential information. It addresses the vanishing gradient problem and enables longer-term information storage compared to traditional RNNs, allowing for continuous learning over multiple time steps and bidirectional propagation. LSTMs employ gate cells to manage information flow, making individual decisions to open or close gates and outperforming regular RNNs in tasks requiring long-term memory.

The CNN-LSTM structure, which merges Convolutional Neural Network (CNN) layers for extracting features with Long Short-Term Memory (LSTM) for predicting sequences, tackles difficulties in tasks such as image captioning.By utilizing pre-trained CNNs to extract image features, this architecture tackles the issue of generating descriptive captions for complex images effectively.

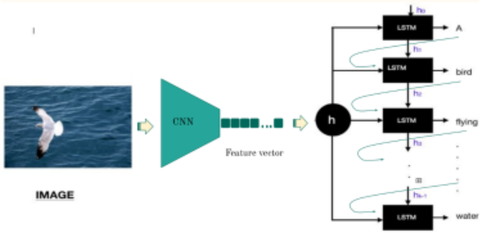


Figure 2. LSTM Network

C. *Encoders: Convolutional Models*

This section covers three encoders. These encoders' descriptions are organized into three subsections. Subsections 1, 2, and 3 address the encoders Resnet50, VGG16, and Inception-V3.

1) Resnet50: A powerhouse in image captioning, this deep convolutional neural network architecture utilizes skip connections to learn residual mappings effectively. With its VGG-style design and up to 50 layers, it captures hierarchical visual patterns efficiently. As an encoder, ResNet50 extracts salient visual features, enabling the generation of rich image descriptions when integrated into captioning models.

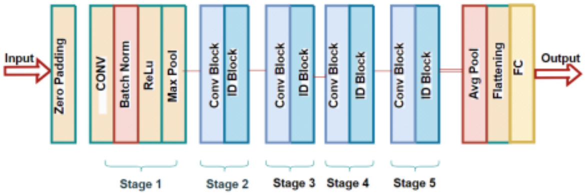


Figure 3. Resnet50 Architecture

2) VGG16: VGG16 is a convolutional neural organize engineering created by the Visual Geometry Bunch (VGG) at the College of Oxford. It is made up of 16 layers, comprising convolutional, pooling, and completely connected layers. In spite of its effortlessness, VGG16 has illustrated amazing execution on a assortment of computer vision applications, counting picture categorization and protest acknowledgment. It is commonly utilized as a highlight extractor in exchange learning and as a benchmark for comparing diverse convolutional neural organize structures.

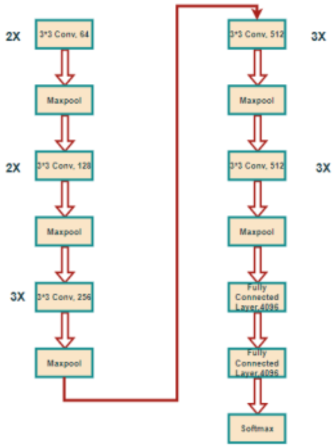


Figure 4. VGG16 Architecture layers

3) InceptionV3: The InceptionV3 is a sophisticated convolutional neural network design that functions as an encoder to extract visual data from images for captioning. It uses inception modules and parallel convolutional filters to efficiently capture multi-scale spatial patterns. The deep stack of these modules, interleaved with pooling layers, encodes high-level semantic visual concepts into a rich feature representation. When used as the encoder in an image captioning model, InceptionV3 provides a robust visual encoding that can be effectively combined with language modeling components to generate descriptive image captions.

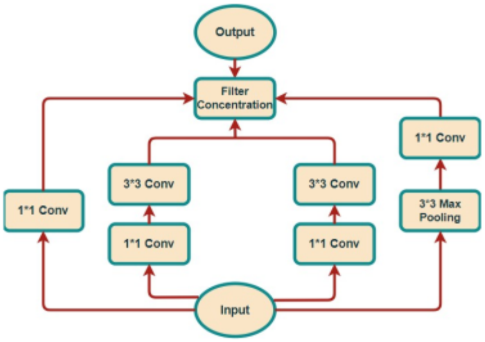


Figure 5. InceptionV3 Architecture

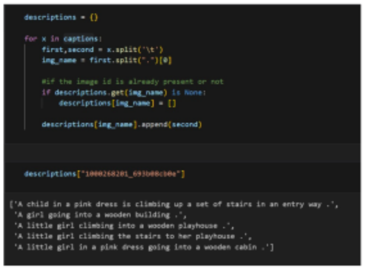
IV. Experiment

A. Dataset Preprocessing

The Flickr8k dataset is the cornerstone of image image research, providing a rich collection of 8,000 images from various groups on the Flickr website. Each image in this dataset is accompanied by five separate captions, adding depth and versatility to the process of training image image models. This amount of captions reduces overfitting and ensures that models generalize well to unseen data.

To streamline the process of model creation and assessment, the dataset is partitioned into three sections: a training subset comprising 6000 images, a development subset with 1000 images, and a test subset containing an additional 1000 images.This partitioning strategy allows researchers to efficiently train, validate, and test their models, ensuring robust performance in real-world scenarios.

First, we loaded the file containing image descriptions and their corresponding IDs. Subsequently, we established a dictionary that associates each photo ID with a corresponding list of text. We cleaned the text by removing punctuation, converting to lowercase, eliminating stop words, and removing tokens with digits to reduce vocabulary size. Next, we created a vocabulary of unique words from all descriptions. Finally, for each training dataset image description, we added ”startseq” at the start and ”endseq” at the end to signify sequence start and end, respectively.



```
descriptions = {}

for x in captions:
    first,second = x.split('\t')
    img_name = first.split(".")[0]

    #if the image id is already present or not
    if descriptions.get(img_name) is None:
        descriptions[img_name] = []

    descriptions[img_name].append(second)
```

```
descriptions["1000268201_693b08cb0e"]
```

```
['A child in a pink dress is climbing up a set of stairs in an entry way .',
 'A girl going into a wooden building .',
 'A little girl climbing into a wooden playhouse .',
 'A little girl climbing the stairs to her playhouse .',
 'A little girl in a pink dress going into a wooden cabin .']
```

Figure 6. Flicker Dataset Python File

B. Feature Extraction

In our consider, photos are utilized as input to the decoder arrange. To prepare the decoder, picture information must be given as fixed-size vectors. Each picture is changed over to a fixed-size vector, which is at that point bolstered into the RNN. We utilize exchange learning to extricate qualities from photos. We utilized pre-trained models with weights determined from comparative information. We utilized pre-trained models to compute picture highlights and spare them in a record. We utilized these characteristics to construct a neural organize to translate the dataset's photos.

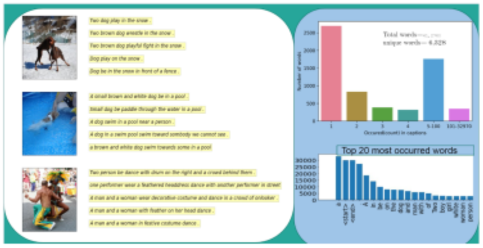


Figure 7. Feature Extraction Dataset

C. Experiment setting

In this project, we compare three pretrained CNN models - VGG16, ResNet50, and InceptionV3 - to assess their performance in automated image processing tasks. We acquire the models from standard libraries like TensorFlow or PyTorch and prepare a diverse dataset encompassing various visual complexities.

After preprocessing and partitioning the dataset into training, validation, and testing sets, we employ transfer learning techniques to fine-tune the pretrained models on our task of automated image captioning. Evaluation criteria like the BLEU score and qualitative examination of generated captions provide information about each model's performance. This systematic comparison aims to identify the strengths and weaknesses of these models for automated image captioning.

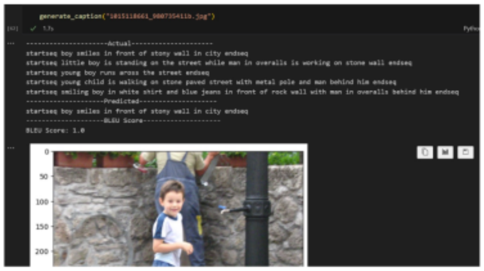
D. BLEU Score

Taking after the era of captions from extricated features, the consequent stage includes surveying the accuracy of the produced captions against the given dataset. Our evaluation utilizes the BLEU score measurements as a gauge to quantify the constancy of our created captions. BLEU score facilitates the examination of the content quality delivered by the Machine Learning show, speaking to one of the earliest metrics to show a solid relationship with human judgments.

The BLEU score ranges between 0 and 1, serving as an indicator of the relevance of machine translation to the actual description. A score of zero signifies no relevance between the machine translation and the reference description, whereas a score of 1 indicates perfect equivalence.

To compute the BLEU score, we generated captions for all images in the test set and used them as candidate sentences.

Each candidate sentence was compared with five reference sentences provided by humans, and a BLEU score was calculated for each comparison. The average BLEU score across all candidate-reference pairs served as an indicator of the accuracy of the generated captions.



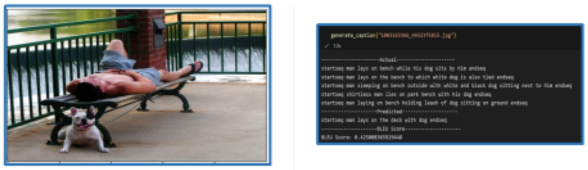
```
generate_caption("1015118661_980735411b.jpg")
```

```
---------------------Actual---------------------
startseq boy smiles in front of stony wall in city endseq
startseq little boy is standing on the street while man in overalls is working on stone wall endseq
startseq young boy runs across the street endseq
startseq young child is walking on stone paved street with metal pole and man behind him endseq
startseq smiling boy in white shirt and blue jeans in front of rock wall with man in overalls behind him endseq
--------------------Predicted--------------------
startseq boy smiles in front of stony wall in city endseq
--------------------BLEU Score--------------------
BLEU Score: 1.0
```

Figure 8. BLEU Score

V. Comparative Analysis of Each Encoder

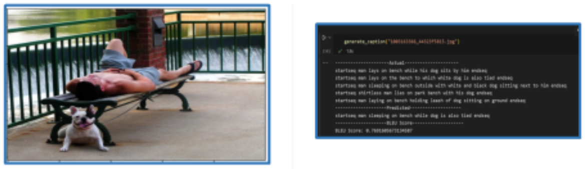
In this part, we compared the accuracy of three encoders: Resnet50, VGG16, and InceptionV3.



```
generate_caption("1003163366_44323f5815.jpg")
```

```
---------------------Actual---------------------
startseq man lays on bench while his dog sits by him endseq
startseq man lays on the bench to which white dog is also tied endseq
startseq man sleeping on bench outside with white and black dog sitting next to him endseq
startseq shirtless man lies on park bench with his dog endseq
startseq man laying on bench holding leash of dog sitting on ground endseq
--------------------Predicted--------------------
startseq man lays on the deck with dog endseq
--------------------BLEU Score--------------------
BLEU Score: 0.425[illegible]
```

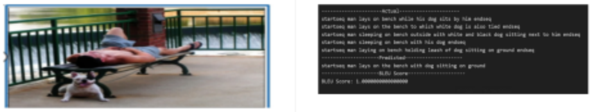
Figure 9. Resnet50



```
generate_caption("1003163366_44323f5815.jpg")
```

```
---------------------Actual---------------------
startseq man lays on bench while his dog sits by him endseq
startseq man lays on the bench to which white dog is also tied endseq
startseq man sleeping on bench outside with white and black dog sitting next to him endseq
startseq shirtless man lies on park bench with his dog endseq
startseq man laying on bench holding leash of dog sitting on ground endseq
--------------------Predicted--------------------
startseq man sleeping on bench while dog is also tied endseq
--------------------BLEU Score--------------------
BLEU Score: 0.7[illegible]
```

Figure 10. VGG-16



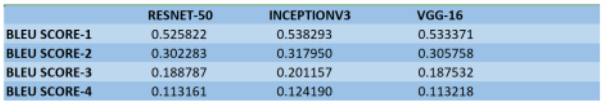
```
---------------------Actual---------------------
startseq man lays on bench while his dog sits by him endseq
startseq man lays on the bench to which white dog is also tied endseq
startseq man sleeping on bench outside with white and black dog sitting next to him endseq
startseq man sleeping on bench with his dog endseq
startseq man laying on bench holding leash of dog sitting on ground endseq
--------------------Predicted--------------------
startseq man lays on the bench with dog sitting on ground
--------------------BLEU Score--------------------
BLEU Score: 1.0000000000000000
```

Figure 11. InceptionV3

VI. Results and Discussion

In the analysis of caption generation performance on the Flickr8k dataset, a total of 810 test images were evaluated. The performance of three different models, namely InceptionV3, ResNet50, and VGG16, was assessed based on the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. These scores are metrics commonly used in natural language processing tasks to evaluate the similarity between generated captions and human reference captions. The results of the evaluation revealed that the InceptionV3 model consistently outperformed the other models across all BLEU scores. This indicates that the captions generated by the InceptionV3 model were closer in resemblance to the reference captions provided by human annotators.

Further analysis showed that the ResNet50 and VGG16 models yielded relatively lower BLEU scores compared to InceptionV3. These lower scores suggest that the captions generated by ResNet50 and VGG16 were less accurate and less similar to the reference captions. This disparity in performance could be attributed to differences in the architecture and features extracted by each model. InceptionV3, known for its effectiveness in image classification tasks, might have better captured the semantic information of the images, leading to more accurate caption generation. On the other hand, ResNet50 and VGG16, while still powerful models, might have faced challenges in capturing finer details or contextual nuances present in the images, resulting in comparatively lower-quality captions. Overall, the analysis highlights the importance of model architecture and feature representation in caption generation tasks, with InceptionV3 emerging as the superior performer in this study.



	RESNET-50	INCEPTIONV3	VGG-16
BLEU SCORE-1	0.525822	0.538293	0.533371
BLEU SCORE-2	0.302283	0.317950	0.305758
BLEU SCORE-3	0.188787	0.201157	0.187532
BLEU SCORE-4	0.113161	0.124190	0.113218

Figure 12. BLEU SCORES

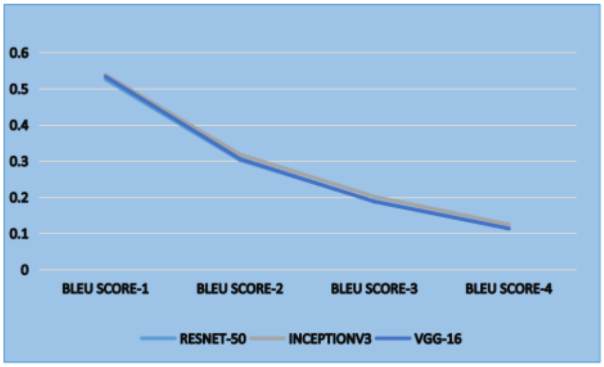


Figure 13. Comparison of BLEU Scores

VII. Conclusion and Future Work

In our evaluation of captions across 810 test images from the Flickr8k dataset, InceptionV3 emerged as the top performer. Its captions consistently achieved the highest BLEU scores, indicating superior accuracy compared to ResNet50 and VGG16. Conversely, the latter models yielded lower BLEU scores, suggesting less precise captions.

The proposed work achieved promising results but was limited by insufficient computational power for fine-tuning hyperparameters like batch size and learning rates. With only 8000 images, the dataset lacked the scale necessary for optimal model training; utilizing larger datasets such as Flickr30k or MS-COCO could improve accuracy and language diversity, but would necessitate high computational resources due to increased training time. Despite these limitations, the study represents a small contribution to a broader research field with ample opportunities for further exploration.

From Pixels to Words: Tracing the Journey of Automated Image Captioning

ORIGINALITY REPORT

4%	2%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Santos Kumar Baliarsingh, Prabhu Prasad Dev, Ashutosh Bhoi, Jasaswi Prasad Mohanty, Jhalak Hota. "Dynamic Population-Based Jaya Algorithm for Genetic Biomarker Classification", 2023 OITS International Conference on Information Technology (OCIT), 2023 Publication	2%
2	Jiabin Luo, Rongzhen Luo. "Research on Image Recognition based on Reinforcement Learning", 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2023 Publication	<1%
3	Ramy Monir, Daniel Kostrzewa, Dariusz Mrozek. "Singing Voice Detection: A Survey", Entropy, 2022 Publication	<1%
4	Sahil Takkar, Anshul Jain, Piyush Adlakha. "Comparative Study of Different Image	<1%

	Captioning Models", 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021 Publication	
5	export.arxiv.org Internet Source	<1%
6	repository.bilkent.edu.tr Internet Source	<1%
7	www.mdpi.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		