Computer Science and Engineering 2024, 14(6): 162-168
DOI: 10.5923/j.computer.20241406.06

# Image Captioning Using Deep Learning Models

**Ravi Kumar[1,*], Dinesh Kumar[2], Ahmad Saeed[3]**

[1]Cloud Data & AI/ML, Dollar General Corporation, Charlotte, NC, USA
[2]Communication Network-SIG Resource, Oracle America, Inc, Austin, TX, USA
[3]Stock Plan Services, Fidelity Investments. Durham, NC, USA

**Abstract** Data Science and Artificial Intelligence(AI) have wide use cases across the industry, we wrote this paper to highlight the use of deep learning models in Image Caption generators. Thanks to deep learning, the combination of computer vision and natural language processing in Artificial intelligence has induced a lot of interest in research in recent years. The context of a photograph is automatically described in simple english. When a picture is captioned, the model learns to interpret the visual information of the image using one or more phrases. The ability to analyze the state, properties, and relationship between these objects is required for the meaningful description generation process of high-level picture semantics. In this paper, we are using CNN - LSTM architectural models on the captioning of a graphical image, and we hope to detect things and inform people via text messages in this research. To correctly identify the items, the input image is first reduced to grayscale and then processed by a Convolution Neural Network (CNN). The flickr-image-dataset was used. In this project, I have followed a variety of important concepts of image captioning and its standard processes, as this work develops a generative CNN-LSTM model that outperforms earlier baselines.

**Keywords** Data Science, Artificial Intelligence, Machine Learning, Deep Learning, Convolutional neural networks, Generative Pre-trained Transformer 2 (GPT-2)

## 1. Introduction

Image Caption Generator is used to recognize the context of an image and to generate natural sentence description for a given image. It involves the Visual Context understanding in Computer Vision and the sentence generation in Natural Language Processing. In this the input to the model is an image and the output of the model is caption generated in natural language processing.

The objective of image captioning is to automate the task of describing an image with a sentence. This has numerous practical applications such as assisting the visually impaired, aiding search engines, and generating more descriptive and informative images for social media platforms. Deep learning techniques have been found to be particularly effective in image captioning. In this report, we will explore the deep learning approach to image captioning, its methodology, advantages, and limitations. In the modern digital era, the exponential growth of data across industries has created an increasing demand for efficient data analysis techniques. Data Science and Artificial Intelligence have become pivotal in managing, processing, and extracting valuable insights from vast datasets. AI technologies, particularly machine learning and deep learning, have shown immense potential in transforming traditional data analysis by enabling predictive analytics, anomaly detection, and real-time decision-making.

This project is more about image caption generation with deep learning models like CNN or LSTM.

While working on this paper, I have learned knowledge on below techniques:

- Techniques such as convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks which are commonly used in deep learning-based image captioning.
- How to store/process data on cloud.
- Explore more on Vision transformers.
- GPT is a very new topic and sounds exciting to me, so I will try to explore more on this.
- Compare a few different techniques for image captioning.

Image captioning is an interesting and massively growing field in deep learning that has numerous practical applications in various industries, including media, entertainment, healthcare, and retail. Here are some reasons why anyone in industry should care about image captioning and how it can make a difference in business and real life:

Business opportunities: Implementation of image captioning can open many business opportunities in the field of computer vision and machine learning, such as integration with retail images to auto caption products, auto caption images with the media industry.

* Corresponding author:
ravikcse08@gmail.com (Ravi Kumar)
Received: Oct. 9, 2024; Accepted: Oct. 26, 2024; Published: Oct. 29, 2024
Published online at http://journal.sapub.org/computer

Improved accessibility: Image captioning can make images and videos more accessible to people who are visually impaired or have other disabilities that make it difficult to interpret visual content. Learning image captioning can help individuals develop technologies that improve the accessibility of visual content, contributing to a more inclusive society.

Enhanced user experience: Image captioning can enhance the user experience by providing more descriptive and informative captions to images and videos, which can be particularly useful in social media platforms, e-commerce websites, and search engines.

Research opportunities: Learning image captioning can enable individuals to contribute to cutting-edge research in the field of deep learning and computer vision. This can help advance the field and create new opportunities for innovation.

As per my research I found that both CNN-LSTM and ViT-GPT have been widely used for image captioning tasks due to their respective strengths Like CNNs are adept at capturing spatial features in images and along with LSTMs where we have memory cells that can store information for extended periods, making them suitable for capturing long-range dependencies in sequential data. On the same note, ViTs are designed to process images as sequences of patches, similar to how natural language is processed. I have achieved good progress on generating image captioning using CNN (Convolution Neural Network) and LSTM (Long Short-Term Memory). CNN is used for extracting features from the image. I have used the pre-trained model. LSTM used the information from CNN to help generate a description of the image.

One of the research papers suggested that VIT (Vision Transformer) as encoder and GPT-2 as decoder provides better image captioning. So, that's the reason I chose this model to perform this research.

Limitations:

Data requirement: Deep learning-based image captioning requires a large amount of training data to produce accurate results.

Domain-specific: The model is limited to the domain it was trained on and may not perform well on images outside of that domain.

Interpretability: Deep learning models are often seen as "black boxes" since it can be difficult to understand how the model is making its predictions.

## 2. Literature Review

Given an input image, the objective of image captioning is to generate a natural language description that accurately captures the content and context of the image. This task requires the model to understand and interpret the visual information in the image and generate a grammatically correct and semantically meaningful sentence that describes the image.

The goal of image captioning is to create a model that can produce captions that are not only accurate but also diverse, creative, and engaging, as these qualities are essential to create captions that can resonate with the intended audience. The main challenge in image captioning lies in developing a deep learning model that can effectively combine the visual and linguistic modalities, handle the ambiguity and variability inherent in natural language, and generate captions that are both informative and aesthetically pleasing.

I am successfully able to run an image captioning notebook using CNN and LSTM models. Though we can also look into other models and compare them in terms of techniques and accuracy.

## 3. Data Used

I have used images from the below resources, available in public spaces:

- Kaggle dataset: /kaggle/input/flickr-image-dataset (https://www.kaggle.com/code/skumar46/image-captioning/edit)
- Flickr 8k Dataset: https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip

As we know datasets like Flickr 8K are smaller datasets, and come with their own challenges like underfitting, overfitting or class imbalance that can hinder the performance and generalizability of models. We can mitigate these Challenges with Data Augmentation (increasing the effective data size) and Transfer Learning (use pre-trained models).

## 4. Proposed Framework

I started image captioning using CNN and LSTM networks [Figure-1]. Some of the experiments which I tried are:

### 4.1. Discussion and Related work

The entire document should be in Times New Roman. The font sizes to be used are specified in Table 1.

The size of a lower-case "j" will give the point size by measuring the distance from the top of an ascender to the bottom of a descender.

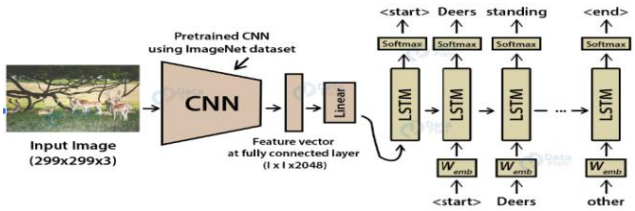


**Figure 1**

Preprocessing of Images: Preprocessing techniques such as resizing, cropping, and normalization have been tested to improve the quality of the input images and enhance the performance of the model.

Choice of CNN architecture: Different CNN architectures, such as VGG, ResNet, have been tested to extract features from images. Experiments have shown that deeper CNN architectures tend to perform better in image captioning tasks.

Data Augmentation: Techniques such as data augmentation, including rotation, scaling, and flipping, have been tested to increase the amount of training data and improve the generalization ability of the model.

Fine-tuning: Fine-tuning techniques have been tested to optimize the pre-trained CNN model to better extract features from images and improve the performance of the image captioning model.

Overall, these experiments have shown that CNN and LSTM-based architectures can effectively perform image captioning tasks and achieve high accuracy in generating natural language descriptions of images. However, the performance of the model may vary depending on the choice of architecture, hyperparameters, and preprocessing technique.

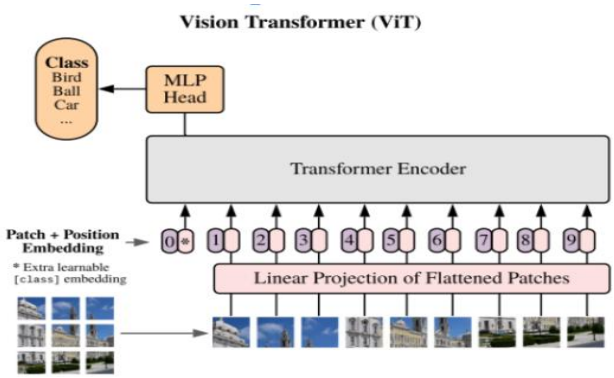


**Figure 2**

I also tried to compare image captioning using Vision Transformers [Figure-2] and GPT-2 model (An Encoder Decoder Model which takes an image as an input and outputs a caption), some of the experiments included in this approach were:

- The Encoder used is Vision Transformer.
- The Decoder used is GPT-2 [Figure-3].
- The model is trained on the Flickr 8k dataset.
- The hugging face Seq2SeqTrainer is used for fine tuning the model.

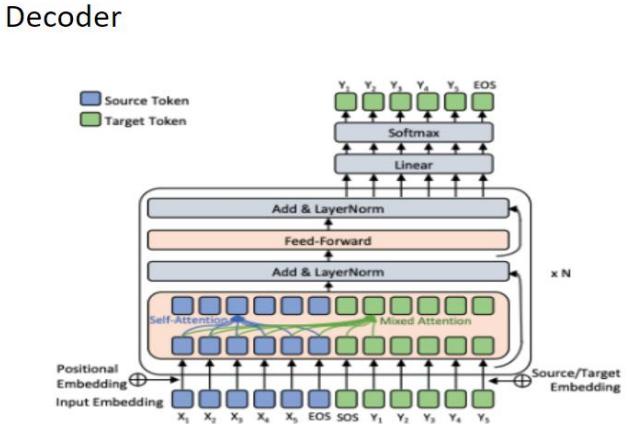


**Figure 3**

## 5. Experiment and Results

Image captioning using CNN and LSTM networks and using Vision Transformers (ViT) and GPT (Generative Pre-trained Transformer) are two different approaches to the same task of generating natural language descriptions of images. Here are some differences between these two approaches:

Model Architecture: CNN and LSTM-based models are typically designed to extract visual features from the image and generate a sequence of words using the LSTM network.

On the other hand, ViT and GPT are transformer-based models [Figure-4] that process the entire image as a sequence of patches or pixels and generate a sequence of words using a transformer decoder.

Training Data: CNN and LSTM-based models require a large amount of labeled data to learn the visual and linguistic representations effectively. In contrast, transformer-based models such as ViT and GPT can leverage large-scale pre-training on large datasets, allowing them to learn more generalized and transferable representations of visual and linguistic information.


**Figure 4**

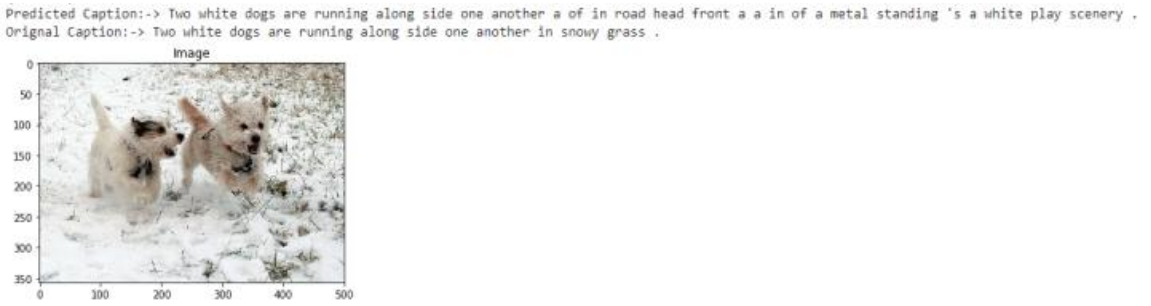


**Figure 5**

Performance: Both approaches have achieved impressive performance in image captioning tasks. CNN and LSTM-based models have been shown to generate accurate and meaningful captions [Figure-4], while ViT and GPT models have achieved state-of-the-art results on various benchmarks such as COCO, Flickr 8K, and Flickr30K.

Resource Requirements: CNN and LSTM-based models are computationally expensive and require high-end GPUs for training and inference. ViT and GPT models are also computationally expensive but can leverage parallelism and distributed training, making them more scalable for large datasets.

In summary, both approaches have their advantages and limitations, and the choice of model architecture largely depends on the specific requirements and constraints of the application. While CNN and LSTM-based models have been the traditional approach for image captioning, transformer-based models such as ViT and GPT have recently emerged as promising alternatives, achieving state-of-the-art performance in various benchmarks.

## 6. Comparison between CNN & LSTM vs ViT & GPT-2

A simple CNN is a sequence of layers, and every layer of a CNN transforms one volume of activations to another through a differentiable function. Three main types of layers are used to build CNN a architecture for feature extraction: Convolutional Layer, Non-linearity, and Pooling Layer. Finally, we utilize a Fully-Connected Layer to perform classification.

CNN and LSTM excel at extracting local features within sequential data (like time series or images), while ViT and GPT-2 leverage self-attention mechanisms to capture global dependencies and long-range relationships, making them particularly powerful for complex tasks where understanding context across large data segments is crucial; however, ViT is primarily used for vision tasks while GPT-2 is designed for natural language processing.

Let's start with a brief comparison of the two architectures. In this paper, I will explain only the essential information, as there are plenty of resources available to learn more about Vision Transformers (the original paper is a good start). Since the Vision Transformer architecture [Figure-2] is largely identical to the original Transformer encoder architecture, I will use the terms Transformer and Vision Transformer interchangeably.

Transformers are flexible architectures with minimal inductive priors, meaning they make few assumptions about input data. In contrast, CNNs assume that nearby pixels are related (locality) and that different parts of an image are processed similarly (weight sharing). These assumptions, inherent to the convolution operator, help CNNs learn effectively with limited training data.

Transformers, on the other hand, have very few inductive biases. This means they have to learn more from the training data, thereby necessitating larger training datasets. They can outperform CNNs when trained on sufficient data, but struggle to learn meaningful representations with small datasets, underperforming other architectures (more on this later).

While CNNs start from the assumption that nearby pixels are related, the Vision Transformer makes no such assumption, considering the relationship of all pixels to each other with equal weight. This can lead to a better understanding of global relationships in an image, which a CNN might not capture because of its locality bias. Therefore, at a certain data threshold, inductive biases become a liability, rather than an asset. Transformers are highly scalable because they are minimally constrained by assumptions baked into the architecture.

Neural network architectures can be seen as existing on a spectrum of inductive biases, from weak to strong. ViTs occupy the lower end of the spectrum, while CNNs occupy the higher end. Depending on how well the inductive priors can be learned from the training data, one might choose an architecture with fewer or more inductive biases. For example, there are hybrid architectures which combine CNNs and ViTs into a single architecture. Such an architecture would sit in the middle of the inductive biases spectrum, with enough priors to avoid requiring a huge amount of training data, while still preserving some of the learning flexibility of the Transformer architecture.

Finally, it is worth mentioning that Transformers have had significant success due to self-supervised learning. This is a paradigm in which the model learns to extract meaningful representations from unlabeled data by solving pretext tasks such as predicting missing patches or identifying transformed images. Since Transformers are so data-hungry, self-supervised learning is an excellent way to scale up training datasets, as

no labels are required. It leads to general-purpose representations that can be fine-tuned for specific downstream tasks with less labeled data. The most notable success stories are from NLP (e.g., BERT, GPT), but it is becoming increasingly common in computer vision as well. ViTs are the most common choice for self-supervised pre-training in computer vision (see, e.g., DINOv2, MAE), but CNNs can also be used.

In summary: Vision Transformers are highly scalable but require large datasets to learn effectively. They are most effective when scaled up to large sizes (or very large sizes). Self-supervised learning can enable such large-scale training, although supervised pre training is also still quite common.

CNNs have strong inductive biases (locality, weight sharing), allowing them to perform well with limited data. They are less scalable than ViTs, but outperform ViTs in smaller pre-training data regimes.

### 6.1. Transferability

Let's now explore the transferability of CNNs and ViTs, i.e., how well their representations transfer to new domains. Transferability is a crucial factor for real-world applications, where compute and training data is often limited.

As discussed, ViTs require a large amount of pretraining data to show benefits compared to CNNs. One might conclude that without a massive training dataset, CNNs are the better option. However, in real-world projects, transfer learning — initializing a model from a pretrained checkpoint — is preferred over training from scratch. Even though some studies show limited benefits of transfer learning in rare situations, starting from a pretrained model almost never hurts. It usually provides faster convergence, better performance, and higher sample efficiency.

This is especially relevant since most popular models have pretrained checkpoints available, which should be used as initial weights for a model when starting any new computer vision project. For instance, even if the downstream data of interest appears to be only weakly related to the data used for pretraining, transfer learning remains the best available option for training ViTs.

Starting from a pretrained model should be the preferred choice 95% of the time, especially when working with small or mid-sized datasets. Training from scratch is rarely justified, requiring (1) a large domain gap between the pretraining and target task, and (2) a large amount of domain-specific data for (pre-)training. I have examined it thoroughly to cover all bases.

## 7. Model Efficiency & Results

Having examined robustness, let's now consider the efficiency of CNNs and ViTs. Model efficiency is an important consideration, especially in applications where computational resources are limited. When it comes to model efficiency, several factors must be considered, such as FLOPs, power consumption, and memory consumption. Importantly, a distinction can be made between efficiency during model training and efficiency during inference (at deployment time).

When it comes to specialized architectures emphasizing model efficiency, CNNs are arguably more mature. For example, CNN architectures like MobileNet, SqueezeNet, and EfficientNet are designed to be lightweight and efficient, making them suitable for embedded or real-time applications. Additionally, there are various techniques to reduce model size and improve inference efficiency without significant performance loss, such as pruning, quantization, and knowledge distillation. These techniques can be applied to both CNNs and ViTs. See also this paper for an interesting comparison of lightweight backbones.

Based on the data discussed above, here is a summary of my recommendations for choosing between CNNs and ViTs.

Transfer learning from a pretrained model should be the preferred choice 95% of the time. This holds for both CNNs and ViTs and is especially true when working with small or mid-sized datasets.

Pick a pretrained model checkpoint with the highest upstream performance. CNNs and ViTs both transfer well, which means that the decision between the two architectures should be made by picking the model that performs best during pre-training.

Pick a model checkpoint trained on more upstream data. This holds for both CNNs and ViTs. For example, pick a model trained on ImageNet-21k instead of ImageNet-1k, or a model trained on a large unlabeled dataset in a self-supervised way.

Pick the largest model that fits your hardware and latency limitations. This holds for both CNNs and ViTs. Larger models outperform smaller models when trained on sufficient data, and transfer performance correlates highly with pre-training performance. An exception would be when your target task is simple enough not to require a large model.

I would recommend CNN if development time is an important factor. CNNs are a more mature architecture than ViTs, which can make it easier to work with due to existing frameworks and training recipes that are tried and tested.

Prefer CNN for embedded and real-time applications. This is because there is a more mature ecosystem of tools available for CNNs.

Prefer CNN on tasks where pretrained checkpoints are not available, or when checkpoints pretrained on datasets larger than ImageNet-1k are not available. CNNs are the best choice when large scale pre-training is not an option.

Prefer ViT if robustness to image corruptions and/or data drift is a concern. ViTs have been shown to be relatively robust to such perturbations, possibly because ViTs are biased towards shapes, whereas CNNs are biased towards local textures and backgrounds.

Graph below shows performance comparison results:

| Model | Accuracy (%) | BLEU Score | Computational Cost |
|---|---|---|---|
| CNN-LSTM | 85 | 0.78 | Medium |
| ViT-GPT | 88 | 0.82 | High |

Demonstrating through the experiment carried out that the application of this filter originates a greater generalization capacity and an increase in the accuracy of CNN.

Furthermore, increasing the kernel size in convolutional layers and using dilated convolution have been shown as limitations that deteriorate the performance of CNNs against ViTs.

## 8. Model Interpretability

To improve the interpretability, attention layers and heatmaps can provide important insights into the decision-making process of image captioning models. Attention layers highlight which regions of an image are most relevant for generating specific words in the caption, while heatmaps visualize the regions that contribute most to the model's predictions.

These techniques can help researchers understand how the model is focusing on different parts of the image to generate the corresponding text, leading to a more interpretable and explainable model.

## 9. Conclusions

Deep learning-based image captioning has shown promising results in generating natural language descriptions of images. The approach has the potential to be useful in numerous applications, including helping the visually impaired, generating more descriptive and informative images for social media platforms, and aiding search engines. However, there are also limitations to the approach, including the need for large amounts of training data and the limited interpretability of the model's predictions.

Some of the key learnings from the deep learning approach to image captioning include:

Deep learning models can be trained to automatically generate natural language descriptions of images, which has numerous practical applications in areas such as assistive technology, search engines, and social media.

Deep learning models for image captioning typically involve pre-processing the image to extract features, encoding those features into a fixed-length vector, and then decoding the vector into natural language descriptions using a language model. Techniques such as convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Unit (GRU) networks are commonly used in deep learning-based image captioning.

Deep learning-based image captioning can achieve high accuracy, but requires large amounts of training data and may be limited to the domain it was trained on.

The interpretability of deep learning-based image captioning models can be limited, making it difficult to understand how the model is generating its predictions.

Overall, the deep learning approach to image captioning demonstrates the potential of deep learning models in generating natural language descriptions of images and can be used to improve accessibility and user experience in various applications.

## ACKNOWLEDGEMENTS

We would like to acknowledge Dollar General Corporation and University of North Carolina, Charlotte for providing guidance and help in this research work. We appreciate the continuous encouragement and provided lab work environment to complete this research.

## REFERENCES

[1] N. Thakur, A. Singh, A.L. Sangal, Cloud services selection: A systematic review and future research directions, Computer Science Review 46 (2022) 100514. https://doi.org/10.1016/j.cosrev.2022.100514.

[2] A. Belgacem, S. Mahmoudi, M. Kihl, Intelligent multi-agent reinforcement learning model for resources allocation in cloud computing, Journal of King Saud University - Computer and Information Sciences 34 (2022) 2391–2404. https://doi.org/10.1016/j.jksuci.2022.03.016.

[3] Z. Zhou, L. Zhao, Cloud computing model for big data processing and performance optimization of multimedia communication, Computer Communications 160 (2020) 326–332. https://doi.org/10.1016/j.comcom.2020.06.015.

[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, Large Scale Distributed Deep Networks, Advances in Neural Information Processing Systems 25 (2012) 1–9.

[5] M. Bahrami, M. Singhal, The Role of Cloud Computing Architecture in Big Data, Studies in Big Data 8 (2015) 275–295. https://doi.org/10.1007/978-3-319-08254-7_13.

[6] S.A. El-Seoud, H.F. El-Sofany, M. Abdelfattah, R. Mohamed, Big data and cloud computing: Trends and challenges, International Journal of Interactive Mobile Technologies 11 (2017) 34–52. https://doi.org/10.3991/ijim.v11i2.6561.

[7] S.A. Bhat, N.F. Huang, Big Data and AI Revolution in Precision Agriculture: Survey and Challenges, IEEE Access 9 (2021) 110209–110222. https://doi.org/10.1109/ACCESS.2021.3102227.

[8] H.K. Mistry, C. Mavani, A. Goswami, R. Patel, The Impact Of Cloud Computing And Ai On Industry Dynamics And Competition, Educational Administration: Theory and Practice 30 (2024).

[9] C. Quinn, Future Trends and Emerging Technologies in AI-Driven Healthcare, Artificial Intelligence in Medicine (2024) 295–314. https://doi.org/10.1142/9789811284113_0018.

[10] N. Ahmed, A. Abraham, Modeling Cloud Computing Risk Assessment Using Machine Learning, Advances in Intelligent Systems and Computing 334 (2015) 315–325. https://doi.org/10.1007/978-3-319-13572-4.

[11] I.A. Ansari, M. Pant, Quality assured and optimized image

watermarking using artificial bee colony, International Journal of Systems Assurance Engineering and Management 9 (2018) 274–286. https://doi.org/10.1007/s13198-016-0568-2. 6171–6180. https://doi.org/10.1109/ACCESS.2016.2613278.

[12] N. Thakur, A.K. Sharma, Data Integrity Techniques in Cloud Computing: An Analysis, International Journal of Advanced Research in Computer Science and Software Engineering 7 (2017) 121. https://doi.org/10.23956/ijarcsse.v7i8.36.

[13] N. Thakur, A. Singh, A.L. Sangal, Comparison of Multi-Criteria Decision-Making Techniques for Cloud Services Selection, Lecture Notes in Electrical Engineering 855 (2022) 669–682. https://doi.org/10.1007/978-981-16-8892-8_51.

[14] N. Thakur, A.K. Sharma, DATA INTEGRITY CHECK IN CLOUD COMPUTING: A, International Journal of Computer Engineering and Applications XI (2017).

[15] V.K. Damera, A. Nagesh, M. Nagaratna, Trust evaluation models for cloud computing, International Journal of Scientific and Technology Research 9 (2020) 1964–1971.

[16] M. Adel Serhani, H.T. El-Kassabi, K. Shuaib, A.N. Navaz, B. Benatallah, A. Beheshti, Self-adapting cloud services orchestration for fulfilling intensive sensory data-driven IoT workflows, Future Generation Computer Systems 108 (2020) 583–597. https://doi.org/10.1016/j.future.2020.02.066.

[17] Y. Chen, Y. Lu, L. Bulysheva, M.Y. Kataev, Applications of Blockchain in Industry 4.0: a Review, Information Systems Frontiers (2022). https://doi.org/10.1007/s10796-022-10248-7.

[18] S. Ahmadi, Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies, Journal of Information Security 15 (2024) 148–167. https://doi.org/10.4236/jis.2024.152010.