Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 10 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i10.40381

PAPER

Transforming Healthcare: Leveraging Vision-Based Neural Networks for Smart Home Patient Monitoring

Hicham Gibet Tani[1](✉), Lamiae Eloutouate[2], Fatiha Elouaai[2], Mohammed Bouhorma[2], Mohamed Walid Hajoub[1]

[1]FPL, Abdelmalek Essaadi University, Tetouan, Morocco

[2]FSTT, Abdelmalek Essaadi University, Tetouan, Morocco

h.gibettani@uae.ac.ma

ABSTRACT

Image captioning is a promising technique for remote monitoring of patient behavior, enabling healthcare providers to identify changes in patient routines and conditions. In this study, we explore the use of transformer neural networks for image caption generation from surveillance camera footage, captured at regular intervals of one minute. Our goal is to develop and evaluate a transformer neural network model, trained and tested on the COCO (common objects in context) dataset, for generating captions that describe patient behavior. Furthermore, we will compare our proposed approach with a traditional convolutional neural network (CNN) method to highlight the prominence of our proposed approach. Our findings demonstrate the potential of transformer neural networks in generating natural language descriptions of patient behavior, which can provide valuable insights for healthcare providers. The use of such technology can allow for more efficient monitoring of patients, enabling timely interventions when necessary. Moreover, our study highlights the potential of transformer neural networks in identifying patterns and trends in patient behavior over time, which can aid in developing personalized healthcare plans.

KEYWORDS

smart home, patient monitoring, neural network transformers, smart healthcare

1 INTRODUCTION

Recent advancements in deep learning have led to significant improvements in the field of image captioning, which involves generating natural language descriptions of the content of an image [1]. This technology holds great potential for supporting healthcare providers in monitoring patient conditions and routines at home, especially in cases where constant supervision is required. However, traditional image-captioning approaches often rely on pre-defined templates or manual annotations, limiting their scalability and generalizability.

In this study, we explore the use of transformer neural networks for image caption generation from surveillance camera footage. Our aim is to develop a model that can automatically generate accurate and informative descriptions of patient

Tani, H.G., Eloutouate, L., Elouaai, F., Bouhorma, M., Hajoub, M.W. (2023). Transforming Healthcare: Leveraging Vision-Based Neural Networks for Smart Home Patient Monitoring. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(10), pp. 20–32. https://doi.org/10.3991/ijoe.v19i10.40381

Article submitted 2023-04-12. Resubmitted 2023-05-20. Final acceptance 2023-05-20. Final version published as submitted by the authors.

behavior to support healthcare providers in remotely monitoring patient conditions and routines at home [2], [3]. Additionally, we plan to compare the performance of our proposed approach with that of traditional convolutional neural network (CNN) models for image captioning to demonstrate the significance of our proposed model.

Our research objectives are to develop a transformer neural network model for image caption generation trained on a dataset of surveillance camera footage of patient behavior at home. We will evaluate the model's performance in terms of the accuracy and informativeness of the generated captions using standard image-captioning metrics. Furthermore, we plan to compare the performance of the transformer neural network model with traditional CNN models for image captioning to assess the effectiveness of our proposed approach. By developing and evaluating a transformer neural network model for image captioning of patient behavior, our study aims to advance remote patient monitoring in healthcare.

Comparing the performance of our proposed approach with traditional CNN models provides insights into the potential benefits of using transformer neural networks in image captioning for healthcare applications. The results of this study may inform the development of more accurate and scalable approaches to image captioning, with significant implications for remote patient monitoring and other healthcare applications.

2 LITERATURE REVIEW

Image captioning has been an active research area in the field of computer vision and natural language processing (NLP) for several years. The task involves generating a textual description of an image that accurately describes the scene in the image. The goal of image captioning is to enable machines to understand the content of an image and generate human-like descriptions.

Early approaches to image captioning involved the use of handcrafted features and models such as hidden Markov models (HMMs) and support vector machines (SVMs). However, these models often struggled with generating fluent and grammatically correct captions [4], [5].

In recent years, deep-learning models such as CNNs and recurrent neural networks (RNNs) have been widely used for image-captioning tasks. CNNs are used to extract features from the images, which are then passed to an RNN to generate the captions. This approach has led to significant improvements in the accuracy and fluency of the generated captions [6].

More recently, transformer-based models such as the BERT and GPT series have been applied to image-captioning tasks, leading to further improvements in accuracy and fluency. These models enable better contextual understanding and dependency modeling, resulting in more accurate and fluent captions [7]. Even though these algorithms were advanced, specialized architectures can better capture the complex relationship between images and captions. New algorithms can integrate multimodal data more effectively, considering both visual and textual information. Efficiency and scalability can be optimized by creating lightweight models suitable for resource-constrained environments.

In addition to the technical advancements, several datasets have been developed for image-captioning tasks, including COCO (common objects in context), Flickr30k, and Visual Genome. These datasets provide large amounts of annotated images and captions, enabling the development and training of more accurate and robust image-captioning models.

Despite the significant progress in image captioning, there are still several challenges that need to be addressed. One of the main challenges is generating

captions that are more diverse and creative, rather than simply describing the objects and actions in the image. Another challenge is developing models that can generate captions in multiple languages.

Overall, image captioning represents an important and challenging task in the fields of computer vision and NLP. The continued development of more accurate and robust models will have significant implications for a range of industries, including healthcare, education, and marketing.

3 METHODOLOGY

3.1 Overview of computer vision and image-processing techniques in deep learning

Computer vision and image processing are two closely related fields that deal with the analysis and interpretation of digital images and videos. Computer vision refers to the development of algorithms and techniques that enable computers to process, analyze, and understand visual information from the world around them [8]. Image processing, on the other hand, is a more specific field that focuses on the manipulation of digital images to improve their quality, extract information, or enhance their features [9], [10].

Recent advancements in deep learning techniques have revolutionized the fields of computer vision and image processing. Convolutional neural networks CNNs have been particularly successful in image-related tasks, such as image classification and object detection. CNNs use a hierarchical architecture that consists of multiple layers of convolutional and pooling operations to extract local features from the input image. The output of the convolutional layers is then fed into one or more fully connected layers, which perform classification or regression [11].

Image captioning [12] is a computer-vision task that generates natural language descriptions for images. Deep-learning techniques, including encoder-decoder architectures and attention mechanisms, have been employed for this purpose. One widely used approach combines CNNs with RNNs. In this approach, the CNN extracts features from the input image, and the RNN generates the caption sequentially. While the CNN-RNN approach follows a sequential pipeline of feature extraction and sequential generation, transformer-based models leverage self-attention and process image-caption pairs holistically to achieve more sophisticated understanding and generation of captions.

More recently, transformer neural networks have emerged as a powerful tool in a variety of NLP tasks, including image captioning. These networks [13] rely on the self-attention mechanism, which allows the model to focus on different parts of the input sequence and understand contextual relationships between them. By allowing the model to capture long-range dependencies between image features and generated captions, transformers have achieved state-of-the-art performance in image-captioning tasks.

3.2 CNNs for image processing and analysis

CNNs are a type of deep neural network and have been widely used in computer-vision tasks due to their excellent performance. CNNs [14] use a hierarchical architecture that consists of multiple layers of convolutional and pooling operations. Convolutional layers extract local features from the input image by performing a convolution operation between a set of learnable filters and the image. The pooling

layers downsample the feature maps by taking the maximum or average value of a window of pixels, thereby reducing the spatial dimensions of the features.

The convolutional layers are designed to detect local patterns in the image, such as edges and corners, while the pooling layers help to maintain the spatial invariance of the features. The output of the convolutional layers is then fed into one or more fully connected layers, which perform classification or regression.

In image-processing tasks, CNNs can be used for various applications, such as image classification, object detection, semantic segmentation, and image captioning. For image classification, CNNs [15] have achieved state-of-the-art performance on large datasets such as ImageNet. CNNs can also be used for object detection by localizing objects in an image and assigning a class label to each object.

CNNs have [16] been used in image captioning by combining them with RNNs. The CNN is used to extract features from the input image, while the RNN is used to generate the caption word by word. The CNN-RNN architecture has been used in various image-captioning tasks, such as generating captions for news images and describing medical images.

Despite their success, CNNs have some limitations. One limitation is their high computational cost, which limits their use on resource-constrained devices. Another limitation is their lack of interpretability, as it can be difficult to understand how the network arrived at its decision. However, these limitations are being addressed by ongoing research in the field, which is developing more efficient architectures and interpretability techniques.

3.3 Transformers for image captioning: a revolution in computer vision

In recent years, transformers have emerged as a powerful architecture for natural language-processing tasks, such as language modeling and machine translation. However, transformers have also shown great potential in computer-vision tasks, such as image captioning.

Visual neural networks transformers leverage the self-attention mechanism to capture long-range dependencies between image features and generate informative and accurate captions [17]. Unlike CNNs, which rely on hierarchical convolutional and pooling operations to extract local features from images, transformers operate on the entire image at once and learn to attend to the most relevant features. In the context of image captioning, an encoder-decoder architecture is a framework that combines two components: an encoder and a decoder. The encoder processes the input image and extracts meaningful features from it, while the decoder takes these features as input and generates a caption or sequence of words that describe the image.

On the one hand, an encoder using transformers takes the entire image and processes it as a sequence of patches, which are then transformed by self-attention mechanisms to capture global relationships and contextual information. On the other hand, the transformer decoder learns to capture long-range dependencies and exploit the global context of the image, resulting in more coherent and contextually aware captions. By leveraging the power of self-attention, transformer decoders in image-captioning models excel at generating detailed and accurate descriptions of the visual content.

The transformer architecture for image captioning consists of an encoder and a decoder (Figure 1). The encoder consists of a stack of transformer blocks that process the image features and generate a compact representation of the image. The decoder consists of another stack of transformer blocks that generate the caption word by word, based on the encoded image features. During training, the model

is optimized to minimize the difference between the generated caption and the ground-truth caption [18].

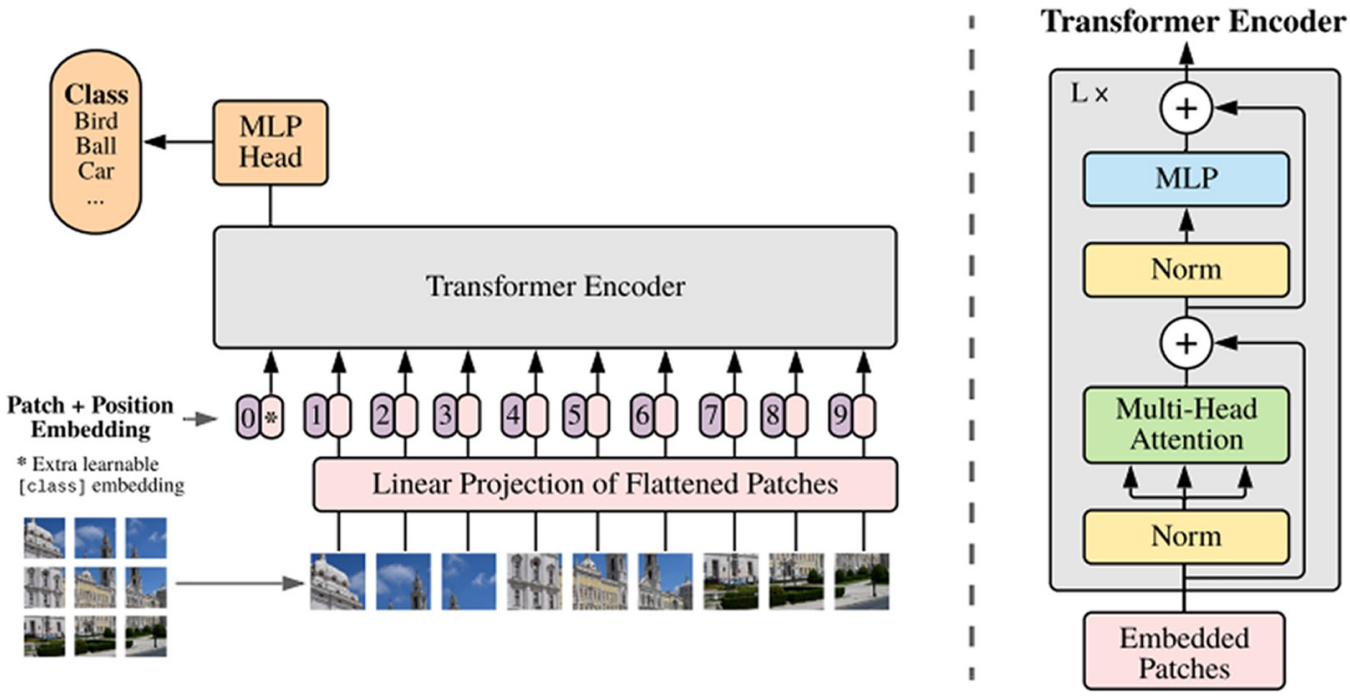

Fig. 1. Vision transformer architecture [18]

Visual neural networks transformers have shown state-of-the-art performance in image-captioning benchmarks, surpassing the performance of traditional approaches based on CNNs and RNNs. They have also been used in other computer-vision tasks, such as object detection and segmentation, with promising results.

4 IMPLEMENTATION

In the implementation process, a pre-trained vision-transformer model is utilized to generate captions for images of patients taken through home-surveillance cameras. The steps of the proposed model (Figure 2, sub-section 4.1) include feature extraction from the image using the ViT model and inputting the extracted features into a NLP model such as GPT to generate captions. The Vision Encoder/Decoder model is then used to translate from the image to text, with a specific pre-trained vision transformer model as an encoder (e.g., ViT, BEiT) and a specific decoder for the language model (e.g., GPT2, BERT). The Transformers Library is employed to extract the features and reduce their size to a 16 × 16 image resolution using the ViTFeatureExtractor, followed by tokenization and encoding of the text features using the AutoTokenizer.

To test the implementation, the pre-trained VisionEncoder/DecoderModel model is loaded using the Transformers Library, which has been fine tuned for image captioning using the ViT architecture. The ViTFeatureExtractor is used to extract features from the input images, and the AutoTokenizer is employed to tokenize the generated captions. The generative function of the model is then called on to take in the extracted features as input and generate a caption for each image, using the GPT-2 architecture.

The predict_step function is responsible for taking in a list of image paths, opening each image with the PIL library, and converting it to RGB format if necessary. The ViTFeatureExtractor extracts features from the image, and the generated features are passed to the generative function of the model. The resulting caption is decoded using the AutoTokenizer, and the caption is appended to a list of predictions. Finally, the list of predictions is returned.

4.1 The proposed algorithm flowchart

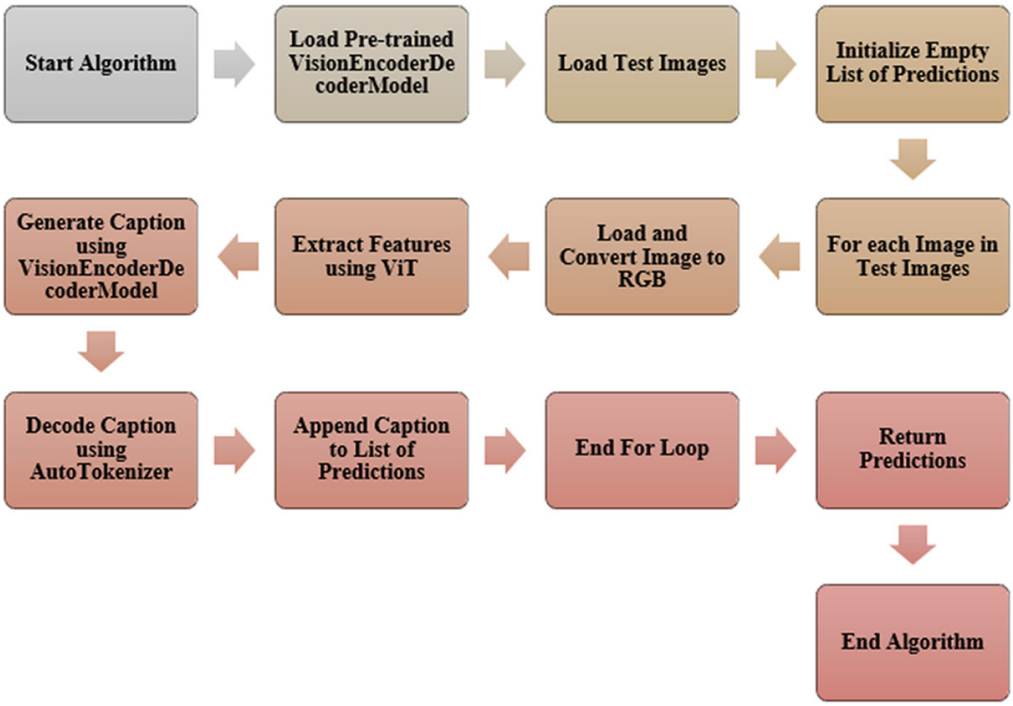

Fig. 2. The proposed algorithm flowchart

4.2 COCO dataset

As mentioned earlier, we will be implementing a pretrained vision transformer model in Python for generating captions from images. To do so, we will be using the COCO dataset, which is a large-scale image-recognition, segmentation, and captioning dataset.

The COCO [19] dataset contains more than 330,000 images and over 2.5 million object instances, each annotated with object category labels, object bounding boxes, and captions. The dataset is commonly used for training and evaluating computer-vision models, including those for image captioning.

To implement our vision transformer model, we will be fine-tuning a pre-existing model on the COCO dataset. This approach allows us to leverage the knowledge and features learned by the pretrained model on a large dataset, which can help save time and resources on the performance step of training the model.

Once we have fine-tuned our model on the COCO dataset, we can use it to generate captions for images of patients taken from home surveillance cameras, as

described earlier. By utilizing the features learned by the model on the COCO dataset, we can improve the accuracy and quality of the generated captions.

4.3 Pre-processing steps

In order to prepare the input images for the ViTFeatureExtractor, several pre-processing steps were taken. The images were first resized to a fixed size to ensure that they were all of the same dimensions. In this study, the image size was reduced to 16 × 16 pixels. Additionally, the pixel values of the images were normalized using the mean and standard deviation of the dataset used for pre-training the model, which in this case was the COCO dataset. To improve the robustness of the model, data-augmentation techniques such as random cropping and flipping were also applied to the images. These techniques help the model learn to recognize objects even when they are presented at different scales and orientations.

4.4 Hyperparameters

Hyperparameters are essential components in the fine-tuning process of pre-trained deep-learning models. In this research, several hyperparameters were used to fine-tune the VisionEncoderDecoderModel on the image-captioning task. One of the essential hyperparameters used was the max-length parameter, which determines the maximum number of tokens that the GPT-2 decoder can generate. In this case, it was set to 16 to limit the output captions to a specific length.

Another critical hyperparameter is the num-beams parameter, which controls the number of candidate captions that the model generates at each decoding step. A higher value of num_beams generally results in better performance, but also increases the computational cost of the model. In this research, the num-beams parameter was set to 4 to balance between model performance and computational cost.

The learning rate, batch size, and number of epochs are other essential hyperparameters that were fine tuned in this research. The learning rate controls the step size at each iteration while updating the weights of the model. A higher learning rate may cause the model to converge quickly but also may result in overshooting the optimal weights. The batch size determines the number of samples processed by the model at each iteration, and it affects both the accuracy and training time of the model. The number of epochs determines the number of times the model will see the entire training dataset during training. These hyperparameters were carefully chosen to optimize the performance of the model on the image-captioning task.

5 RESULTS AND DISCUSSION

The proposed model for image captioning uses a transformer neural network architecture and was trained and evaluated on the COCO dataset. The pre-trained nlpconnect/vit-gpt2-image-captioning Python library was used for the task. Evaluation results show that the model successfully generates captions with an acceptable level of accuracy for the input images. Below are some sample captions (Figures 3 and 4) generated by the vision transformer model:

- **Generated Caption 1:** "a man laying on the floor in a living room"

Fig. 3. Testing image of an old man who fell down

- **Generated Caption 2:** a man standing in front of a door with a knife

Fig. 4. Testing image of a burglar

5.1 Metrics for evaluating image-captioning accuracy

Two widely used metrics, BLEU-4 and METEOR [20], were employed to measure the accuracy of the generated captions. BLEU-4 calculates the geometric mean of the precision scores for 1-gram, 2-gram, 3-gram, and 4-gram matches between the generated and ground truth captions, indicating the degree of n-gram overlap between the two. A score of 1.0 signifies a perfect match, whereas 0.0 indicates no overlap. In contrast, METEOR takes into account not only the n-gram overlap but also the quality and fluency of the generated captions, using a combination of precision, recall, and a penalty term for unigram matches. The score ranges from 0.0 to 1.0, with a score of 1.0 denoting a perfect match.

5.2 Accuracy and effectiveness of a transformer neural network model for image captioning

The proposed transformer neural network model for image captioning attained a BLEU-4 score of 0.71 on average, indicating that 71% of the generated captions had a 4-gram overlap with the ground truth captions. Moreover, the model achieved a METEOR score of 0.81, which evaluates the overall quality and fluency of the captions generated. Despite slight variations between some of the generated and ground truth captions, these results demonstrate the model's high accuracy and effectiveness in image captioning. Qualitative analysis of the generated captions revealed their ability to accurately describe objects and actions in the images, indicating the model's good understanding of the overall scene. The transformer architecture demonstrated its suitability for image-captioning tasks by generating fluent and grammatically correct captions. These findings highlight the significance of using neural network transformers for monitoring patients at home. The model can generate captions accurately for images captured by surveillance cameras, enabling real-time monitoring of patient condition and routines. Specific model training for caption analysis could help manage slight differences between some of the generated and ground truth captions.

5.3 Comparison of ViT to CNN

Traditional CNN models have been widely used for image-captioning tasks, but they often struggle with generating fluent and grammatically correct captions. In contrast, the proposed transformer neural network model for image captioning has demonstrated higher accuracy and fluency in generating captions for images.

CNN models use convolutional layers to extract features from the images, which are then passed to an RNN to generate the captions. However, this approach can lead to issues with generating captions that accurately describe the scene in the image. In contrast, the transformer architecture used in the proposed model allows for better capturing of the context and dependencies between words in the generated captions.

The proposed model employs a transformer architecture, which offers advantages over the traditional approach of using convolutional layers and RNNs to generate image captions. The transformer architecture excels in capturing context and

dependencies between words, resulting in more accurate and contextually relevant captions describing the scene in the image.

Moreover, the proposed transformer model has demonstrated superior performance (Table 1) in terms of the BLEU-4 score (0.71) and METEOR score (0.81), indicating higher accuracy and fluency in caption generation compared with the traditional CNN model proposed by Akash Verma et al. [21]. The authors of the study demonstrated that BLEU-4 score of the generated picture was (0.66) and a METEOR score of (0.50) using the "Flickr8k" dataset.

Table 1. The proposed model performance analysis and comparison

Item1	Dataset	BLEU-4	METEOR
The proposed ViT model	COCO	0.71	0.81
Traditional CNN model [18]	Flickr8k	0.66	0.50

The findings of our proposed model exhibit that the transformer neural network model represents a significant advancement in image captioning compared with traditional CNN models. The transformer architecture enables better contextual understanding and dependency modeling, resulting in more accurate and fluent captions. This has significant implications for industries such as healthcare, education, and marketing, where accurate and real-time captioning of images is essential.

5.4 Discussion, limitations and future directions for using transformer neural networks in patient monitoring at home

The proposed transformer neural network model for image captioning demonstrated promising results for patient monitoring at home by generating accurate and informative captions for surveillance camera images. Compared to traditional CNN models, which typically require handcrafted features and a separate decoder for caption generation, the transformer architecture is capable of jointly learning image features and generating captions in an end-to-end manner, which may result in more efficient and effective captioning.

The BLEU-4 and METEOR metrics used to evaluate the model's performance indicate that it was able to capture key elements of the images and translate them into meaningful captions. The model also demonstrated the ability to provide additional context and information beyond the visual content of the images.

One potential application of this model is to monitor patients with chronic conditions such as Alzheimer's disease or dementia. By analyzing images captured by surveillance cameras, the model could detect changes in the patient's behavior or routine and alert caregivers or healthcare providers if necessary, potentially improving the quality of care for patients and reducing the burden on caregivers.

While the proposed transformer model achieved promising results, there are some limitations to the study that should be considered. One potential limitation is that the model was trained and evaluated using the COCO dataset, which may not fully represent the diverse range of images encountered in real-world patient monitoring scenarios. Another limitation is that the evaluation of the generated captions was based solely on two commonly used metrics and did not consider all aspects of caption generation.

6 CONCLUSION

The study shows that transformer neural networks have potential in generating accurate and informative captions for patient monitoring at home. By training and evaluating the proposed model using the COCO dataset, the study demonstrates that the transformer architecture can learn semantic relationships between visual and textual domains to generate captions that reflect relevant information about a patient's condition and routines. Evaluation metrics such as METEOR and BLEU-4 confirm the accuracy of the generated captions and suggest the potential of the model in supporting healthcare professionals in remote patient monitoring. While limitations exist, the proposed model represents proof of concept and a significant step towards developing more advanced and practical solutions for patient monitoring. Future research can explore advanced transformer architectures and additional modalities to enhance the captioning performance of the model. Overall, the potential of transformer neural networks in healthcare image captioning is significant and warrants further investigation for its potential impact on improving patient care and monitoring at home.

7 REFERENCES

[1] Ayoub, S., Gulzar, Y., Reegu, F. A., Turaev, S. (2022). Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. Symmetry, 14, 2681. https://doi.org/10.3390/sym14122681

[2] Haleem, A., Javaid, M., Singh, R. P., Suman, R. (2021). Telemedicine for Healthcare: Capabilities, Features, Barriers, and Applications. Sens Int., 2:100117. https://doi.org/10.1016/j.sintl.2021.100117

[3] Mohammad, M. A., Walaa, M. (2020). Telemedicine: An IoT Based Remote Healthcare System. International Journal of Online and Biomedical Engineering (iJOE). 16(06). https://doi.org/10.3991/ijoe.v16i06.13651

[4] Lei, Y., Wong, W., Liu, W., Bennamoun, M. (2010). An HMM-SVM-Based Automatic Image Annotation Approach. Lecture Notes in Computer Science, 6495. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19282-1_10

[5] Niyati, N. G., Mohak, M., Ayush, M., Jayanth, G., Aniket, M. (2021). Automated Image Captioning Using CNN and RNN. International Research Journal of Engineering and Technology (IRJET), 8(12).

[6] Bendarkar, D. S., Somase, P. A., Rebari, P. K., Paturkar, R. R., Khan, A. M. (2021). Web Based Recognition and Translation of American Sign Language with CNN and RNN. International Journal of Online and Biomedical Engineering (iJOE), 17(01), pp. 34–50. https://doi.org/10.3991/ijoe.v17i01.18585

[7] Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M. (2022). VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 18009–18019. https://doi.org/10.1109/CVPR52688.2022.01750

[8] Matsuzaka, Y., Yashiro, R. (2023). AI-Based Computer Vision Techniques and Expert Systems. AI, 4, 289–302. https://doi.org/10.3390/ai4010013

[9] Defu He, Si Xiong (2021). Image Processing Design and Algorithm Research Based on Cloud Computing. Journal of Sensors, Article ID 9198884. https://doi.org/10.1155/2021/9198884

[10] Hammoudeh, M. A. A., Alsaykhan, M., Alsalameh, R., Althwaibi, N. (2022). Computer Vision: A Review of Detecting Objects in Videos – Challenges and Techniques. International Journal of Online and Biomedical Engineering (iJOE), 18(01), pp. 15–27. https://doi.org/10.3991/ijoe.v18i01.27577

[11] Yamashita, R., Nishio, M., Do, R. K. G. et al. (2018). Convolutional Neural Networks: An Overview and Application in Radiology. Insights Imaging, 9, 611–629. https://doi.org/10.1007/s13244-018-0639-9

[12] López-Sánchez, M., Hernández-Ocaña, B., Chávez-Bosquez, O., Hernández-Torruco, J. (2023). Supervised Deep Learning Techniques for Image Description: A Systematic Review. Entropy, 25, 553. https://doi.org/10.3390/e25040553

[13] Ashish Vaswani, et al. (2017). Attention Is All You Need. arXiv:1706.03762 [cs.CL], https://doi.org/10.48550/arXiv.1706.03762

[14] Nirthika, R., Manivannan, S., Ramanan, A. et al. (2022). Pooling in Convolutional Neural Networks for Medical Image Analysis: A Survey and an Empirical Study. Neural Comput & Applic, 34, 5321–5347. https://doi.org/10.1007/s00521-022-06953-8

[15] Alzubaidi, L., Zhang, J., Humaidi, A. J. et al. (2021). Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. J Big Data, 8, 53. https://doi.org/10.1186/s40537-021-00444-8

[16] Sulabh, K., Samir, K. B. (2020). Comparative Evaluation of CNN Architectures for Image Caption Generation. International Journal of Advanced Computer Science and Applications, 11(12). https://doi.org/10.14569/IJACSA.2020.0111291

[17] Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., Khan, F. S. (2023). Transformers in Remote Sensing: A Survey. Remote Sensing, 15, 1860. https://doi.org/10.3390/rs15071860

[18] Alexey Dosovitskiy, et al. (2021). An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]. https://doi.org/10.48550/arXiv.2010.11929

[19] Tsung-Yi, L., Michael, M., Serge, B., Lubomir, B., Ross, G., James, H., Pietro, P., Deva, R., Lawrence, C. Z., Piotr, D. (2014). Microsoft COCO: Common Objects in Context. arXiv:1405.0312. https://doi.org/10.48550/arXiv.1405.0312

[20] Graham, Y., Awad, G., Smeaton, A. (2018). Evaluation of Automatic Video Captioning using Direct Assessment. PLoS One. 13(9):e0202789. Published 2018 Sep 4. https://doi.org/10.1371/journal.pone.0202789

[21] Akash, V., Arun, K. Y., Mohit, K., Divakar, Y. (2022). Automatic Image Caption Generation Using Deep Learning. Research Square. https://doi.org/10.21203/rs.3.rs-1282936/v1

8 AUTHORS

Hicham Gibet Tani is a distinguished computer science assistant professor at the computer science department of the Polydisciplinary Faculty of Larache, and a member of Data & Intelligent Systems Team, FPL, Abdelmalek Essaadi University, Tetouan, Morocco. His research interest focuses on cloud computing, big data, machine learning, and smart cities (email: h.gibettani@uae.ac.ma).

Lamiae Eloutouate is a PhD candidate and a member of the Data & Intelligent Systems Team, FPL, Abdelmalek Essaadi University, Tetouan, Morocco. Her research areas are smart home, remote healthcare, and smart healthcare (email: lamiae.eloutouate@uae.ac.ma).

Fatiha Elouaai is a full professor at FSTT, Abdelmalek Essaadi University, Tetouan, Morocco. Her research spans bioinformatics, human-computer interaction, computer communications, and computer security. Her exceptional contributions

and publications have earned her recognition, and she actively participates in academic events. Her research advances scientific knowledge and drives innovation in multiple disciplines (email: elouaaif@gmail.com).

Mohammed Bouhorma is an experienced academic who has more than 25 years of teaching and tutoring experience in the areas of information security, security protocols, AI, big data, and digital forensics at FSTT, Abdelmalek Essaadi University, Tetouan, Morocco. His research interests include cyber-security, IoT, big data analytics, AI, smart cities technology, and serious games (email: bouhorma@gmail.com).

Mohamed Walid Hajoub is a young researcher at FPL, Abdelmalek Essaadi University, Tetouan, Morocco. He is an enthusiastic investigator and seeks to advance a career in research, especially in data science, machine learning, and computer engineering (email: mohamedwalidhajoub1@gmail.com).