International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 991-999
https://doi.org/10.70917/ijcisim-2026-4254
Article

A Neuro-symbolic Framework for Hallucination Detection in Image Captioning Using Knuth–Morris–Pratt Pattern Matching

Haraprasad Naik*[1], Prafulla Kumar Behera[2]

[1]Department of Computer Science and Applications, Utkal University, Bhubaneswar-751004, Odisha, India.
Orcid Id- 0000-0001-8043-3540
*Corresponding author
Email: hnaik.cs@utkaluniversity.ac.in
[2]Department of Computer Science and Applications, Utkal University, Bhubaneswar-751004, Odisha, India.
Orcid Id- 0000-0001-5016-0692
Email: pkbehera.cs@utkaluniversity.ac.in

Abstract: The research on image captioning has improved significantly in terms of describe of a scene appears on an image. This particular task of image caption generation is considered as most vital because of its wide scope multidimensional application towards various societal usages. However, the existing caption generators could not be deployed with confidence which could serve 100% accuracy because most of them suffers from hallucination. That means a large number of caption generator still struggling to achieve the human like performance as far the semantic values or context of the sentence is concerned. We could predict the potentiality of image captioning task that can leverage many other areas of computer science by simply observing the exponentials growth in this topic. In spite of this popularity, the hallucination detection is most prominent issue nowadays that could finetune the performance of the image caption generators. The term hallucination refers to the issue of predicting nonexistent object automatically solely due to the biasness of training data. In other words, when the caption generator automatically predicts a non-existing object from the input image this is called as object hallucination. Although the hallucination are three different types such as object hallucination, spatial hallucination and multimodal hallucination, the object hallucination is the most prominent one in this field. To address this object hallucination problem, we have proposed a neuro-symbolic framework that not only generates the semantically correct caption but also successfully detects the object hallucination by integrating the classic Knuth-Morris-Pratt (KMP) pattern matching algorithm. We have demonstrated this substring matching process by utilizing the KMP algorithm and found a linear polynomial time complexity with in the controlled environment of experimental setup.
Keywords: Hallucination, Pattern Matching, Image Captioning, Object Detection, Visual Language Models

1. Introduction

The research direction of captioning from an image is generally considered as a multimodal topic that contributes in the field of natural language processing and image processing. Now a days, most of the image caption generator primarily based on the transformer and claimed remarkable improvements in object detection from an image. If we delve more into the topic of object detection, we might notice a significant advancement year after year from the initial idea in this domain. Most of the researchers considered the Deep Neural Networks[14] as a part of their visual encoder that could successfully achieve expected benchmark level performance. However, we could notice that there was no such pace in the contribution of caption formation. Therefore, in this work we have proposed a novel framework that not only generate caption but also improves the quality of caption at par with human generated caption by considering the semantic values of each object and their localization. Basically, the image caption generation

models are largely affected with certain issues that appears as the real concern among its researchers. The Hallucination Detection is one of such problem that has emerged recently as primary concern. Hallucination detection in image captioning task gained attention due to its variability nature in its types. Basically, the hallucination could be categorized into numerous categories but we have discussed these three categories: 1. Object Hallucination[7] 2. Spatial Hallucination[15] 3. Multimodal Hallucination[18]. The First category of hallucination appears when the object detection model has not yet detected an object but the caption generator included the object name as an effect of hallucination. The second category appears when the location of the object could not be localized correctly due to insufficient extracted features. Finally, the third category includes the repetitive object name in the desired caption as well as the presence of the former two objects. Although the hallucination detection captivated tremendous attention amongst the researchers, however most of such detection techniques relies on the explainable evaluation framework utilizes the metrics such as CHAIR, FAITHSCORE and ALOHa. In this work, we have proposed an Image Caption Verification Framework by applying the string matching KMP algorithm that not only eliminates the repetitive object names from the caption but also identify object hallucination from the generated caption as discussed above. We have emphasized on the implementation of KMP algorithm because of its linear time complexity of $O(M+N)$ to search an object name from the generated caption. In this proposed Framework, the overall idea could be summarized in following steps:

- An Encoder-Decoder based model would generate an Image Caption as well as returns the object names present in the input image.
- The KMP algorithm would check or verify the presence of object name in the generated caption.

Figure 1 : The sample captions generated from the model

2. Background and Related Works

The hallucination detection with LLMs for the NLP task has made remarkable stride, but in the domain of image captioning yet to explore significantly in terms of research directions. Since the image captioning is a multimodal task and connects the modalities of NLP with Image processing, we have harnessed the concept of transfer learning instead of building our own transformer[19] from the scratch. In this work we have utilized the Dense201 model which is a Convolutional Neural Network (CNN) Model as encoder and the LSTM as the decoder. The principal idea is to produce a list of objects name that are exist in the input image and store in a feature vector. Afterwards, the LSTM would produce a semantically correct caption and also be stored in another vector. During the object detection the CNN has applied an additional layer of multi-headed attention so that the object localization could be perfectly achieved. The KMP pattern matching algorithm is utilized to verify the appearance of all the objects that has been detected earlier are present in the system generated caption. To avoid the unnecessary computation, we have optimally chosen the beam size during the application of beam search.

The primary motivation towards building the transformer based visual language model is to minimize the hyperparameter tuning to reduce the computation. This minimization achieved through the zero-shot generalization. In recent years image captioning gained attention of many researchers from the field of image processing and computer vision. Accordingly, the state-of-the-art visual language models are being proposed which includes BLIP, Llava, GiT etc. All of such models are lacking the qualitative parameters in terms of evaluation of generated caption at par with

human does. That means the semantics of the generated caption still requires more emphasis in terms of further research. Although there exist multiple evaluation metrics which focuses on the linguistic quality of the generated caption. These metrics includes BLEU, ROUGE, SPICE and CIDEr.[14] However, in terms of hallucination detection[7] the outcomes of these metrics may mislead. However from the above listed metrics only SPICE and CIDEr metrics are most often utilized for image captioning. In this paper we have used the metrics such as CHAIR[18], FAITHSCORE[11] and ALOHa[6] which are exclusively proposed to measure hallucination detection. Here the visual transformer such as Llama which was proposed by Meta could also be implemented with nominal hardware configuration or advancement thereon. Specifically, the Llama.cpp has marked significant performance than any of its counterpart exist in cloud environment.

The CHAIR metric is the baseline metric and has two different variants:

$$CHAIR_i = \frac{|HallucinatedObjects|}{|AllpredictedObjects|}$$

and

$$CHAIR_s = \frac{|Sentenceswithhallucinatedobjects|}{|Allsentences|}$$

The former one has been utilized by the researchers to measure hallucination per sentence and other is for each instance. The CHAIR is the acronym for “Caption Hallucination Assessment with Image Relevance”. This metric often utilized as the first and simple measure to detect hallucination.

The FAITHSCORE is another metric which measures the performance of hallucination detector by dividing complete caption into smaller sub captions. From these sub captions the fine-grained atomic facts are extracted to measure the performance. In a same vein, The ALOHa could also be utilized to measure performance of the hallucination detector. This metric utilizes the Hungarian matching algorithm to find semantic similarities between the actual or grounded object from the candidate caption with reference object from the entire caption.

There exists another hallucination evaluation method which is based on polling and answers query of the evaluator in YES/NO format. This metric is known as POPE which is the acronym for Polling-based object Probing Evaluation.

The proposed Framework successfully detects the object hallucination by analyzing the generated caption against the list of objects names. If the detected objects are stored into a vector such as:

and each word of the generated caption is mapped onto another vector such as:

then the KMP pattern matching algorithm detects the object hallucination. In order to evaluate the fine-grained objects along with their context from the image caption, this hallucination detection does not embed each of the word with a number, rather it compares the two vectors as said in the above algorithm to found discrepancies. Here discrepancies tend to state that if a phantom occurs of a nonexistent object into the caption. There exists wu-palmer similarity[15] check mechanism that could evaluate the context of each word for replacement with synsets of WordNet. The M. Lymperation et al.[15] proposed a framework called “HalCECE” that provides semantically meaningful edits to the generated caption. They have focused on the concept of "hypernyms" that denotes the most suitable word in the replacement of the existing one. In the same work they have investigated about the role hallucination, that interconnects between the visual objects and its localization. As we have stated in the above captioned image two dogs running or playing could also be justified with role hallucination detection. Here the KMP algorithm[12] is presented with its pseudocode version for illustration purpose.

Algorithm 1 Knuth–Morris–Pratt (KMP) Algorithm

Require: Caption Text ‘T’ of size n, Object List pattern P of size m

Ensure: The Initial Index where P matches in T

1: Evaluate the Longest Prefix Suffix (LPS) for the object list pattern P

2: x ← 0 {First index of T}

3: y ← 0 {First index of P}

4: while (x < n) do:

5: if P [y] = T [x] then

6: x← x + 1

7: y ← y + 1

8: end if

9: if y = m then

10: declare the matching at index x − y

11: y ← LPS [y − 1]

12: else if x < n and P [y] = T [x] then

13: if y= 0 then

14: y ← LPS [y − 1]

15: else

16: x ← x + 1

17: end if

18: end if

19: end while

Algorithm 2 : The Longest Prefix Suffix (LPS) evaluation

1: LPS [0] ← 0

2: length ← 0

3: x← 1

4: while x < m do

5: if P [x] = P [length] then

6: length ← length + 1

7: LPS[x] ← length

8: x ← x + 1

9: else

10: if length= 0 then

11: length ← LPS[length − 1]

12: else

13: LPS[x] ← 0

14: x ← x + 1

15: end if

16: end if

17: end while

Source: Fast Pattern Matching in Strings, SIAM Journal of Computing, Vol-6, No-2, June 1977

3. The Architecture of Proposed Framework

The proposed framework is composed of two components. The first component is meant for the caption generation and other is to detect hallucination. The proposed caption generation is utilizing the transfer learning using Dense201[13] network for generating caption and the LSTM[24] as the decoder of the caption generator. After successful caption generation, it would create a new vector to store the generated caption and then compare it with the list of objects name to detect object hallucination.

In the proposed model, we have used the built-in **conv2d()** method of TensorFlow as:

conv2d(i,w,strides=[1,1,1,1], padding='SAME')

Where, i is the data of the input image.

These image data are then downsized with feature extraction techniques. This downsized feature map is simply a term usually refers to output of each sub-layers of CNN. There exists another way to visualize the output of the processed image after applying the filter onto the input image. this filter is parameterized using the list of parameters mentioned in the above equation as 'w' and this is the learned weight in a small sliding window. The output of this operation merely depends on the shape of i and w. The image data i has the shape in the vector form as: [None, 28,28,1] means we have an unknown numbers of images each of size 28 x 28 along with one colour channel. Similarly, the weight 'w' has the shape of [5,5,1,32] means the 5x5x1 denoting the size of the window, which is convolve with the pixels of the image. The number 32 in that vector denoting the numbers of feature map. The stride argument is responsible for the spatial movement of the filter across the input image 'i'. Here the value [1,1,1,1] is denoting the filter could be applied to the input image with one pixel interval in each dimension. This corresponds to the full convolution. Finally, the argument 'SAME' is applied to the padding, that means the border of image 'i' are padded in such a manner that the resultant matrix would be of same size.

Following the linear layers of convolution, some activation functions are generally utilized to add non-linearity. The max pooling is utilized in this model as:

max_pool(i,ksize=[1,2,2,1], stride=[1,2,2,1], padding= 'SAME')

where **i** is the input image and ksize denotes the kernel size.

The above expression of max pooling selects the maximum output of a particular region as the filter stride as sliding window. Here the ksize controls the size of the pooling and stride indicates how much slide the pooling window would make. After all such layers, finally the dropout layer needs to be framed as the regularization task. This dropout turns off multiple units randomly by turning their values to zero during training. The architecture and the parameters details could be observed from the following figure:

Model: "densenet201"

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 224, 224, 3)	0	-
zero_padding2d (ZeroPadding2D)	(None, 230, 230, 3)	0	input_layer[0][0]
conv1_conv (Conv2D)	(None, 112, 112, 64)	9,408	zero_padding2d[0…
conv1_bn (BatchNormalizatio…	(None, 112, 112, 64)	256	conv1_conv[0][0]
conv1_relu (Activation)	(None, 112, 112, 64)	0	conv1_bn[0][0]
zero_padding2d_1 (ZeroPadding2D)	(None, 114, 114, 64)	0	conv1_relu[0][0]

Total params: 20,242,984 (77.22 MB)

Trainable params: 20,013,928 (76.35 MB)

Non-trainable params: 229,056 (894.75 KB)

Figure 2: The proposed model with input images having dimension 224 x 224 x 3

The model would generate the two vectors where one vector is meant for the generated caption and another for the list of objects names. The KMP algorithm would then be utilized to check the object hallucination by applying sub string matching techniques. The illustration of string matching could be visualized as follows:

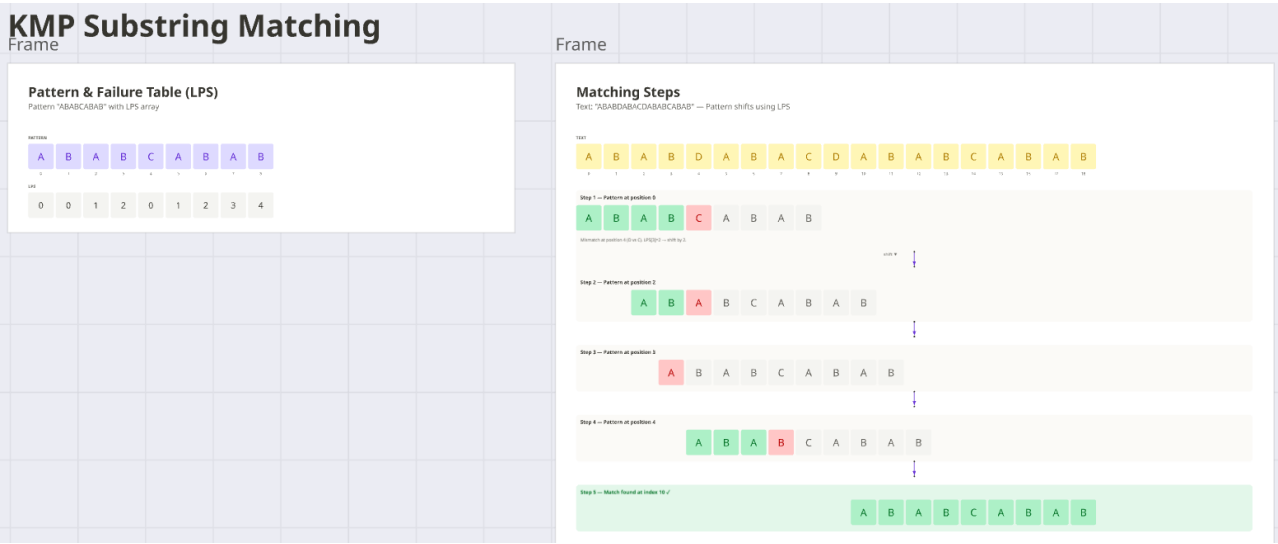

Figure : Sample Text with pattern matching

Here we could observe that the pattern would be checked for matching alongside the original text. This matching would detect the objects name in the original model generated caption. For example, for the first caption of figure-1 would be checked against the object “ball”, that could be illustrated as follows:

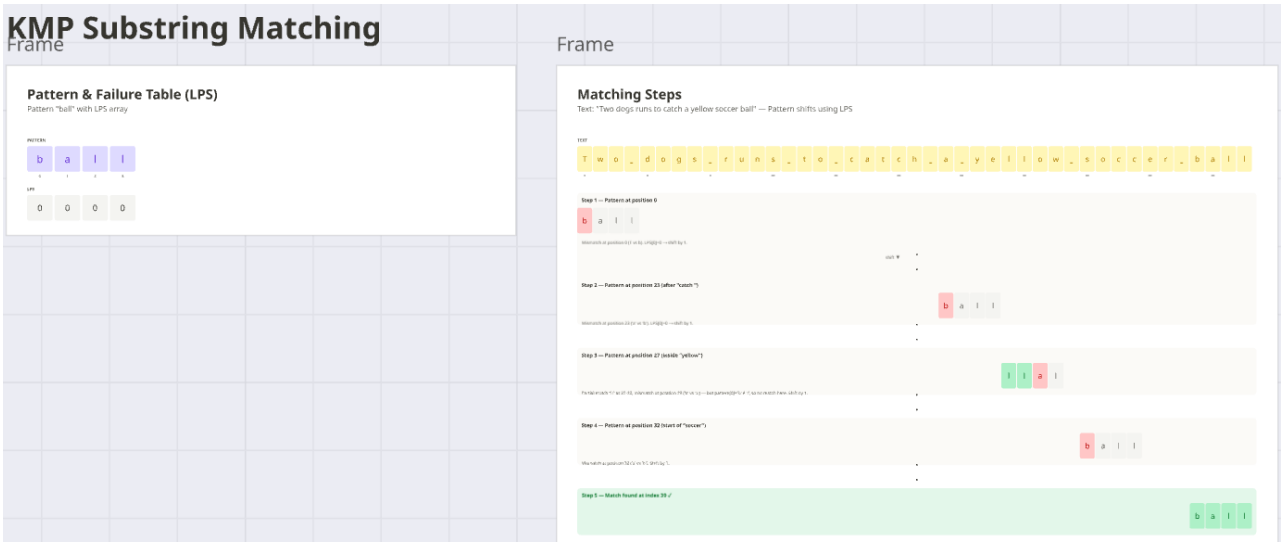

Figure 4: Caption with Object name matching

4. The Experiment and the Dataset

We have used the Flickr 8K dataset to train the model from which the sample captions are generated (illustrated in Figure-1). This dataset has total 8092 images and all are in the jpeg format. Each of these images has at least five human annotated captions which are stored in a separate text file 'caption.txt'. This text file contains approximately more than 40000 human annotated captions. That means at least five captions for each image of the dataset. The training set is comprises of 6000 images along with 1000 images available for each validation and testing set. We have utilized the Kaggle notebook with accelerator 'P100 GPU' for training of the model. The learning rate could be observed by checking the loss curve as illustrated below.

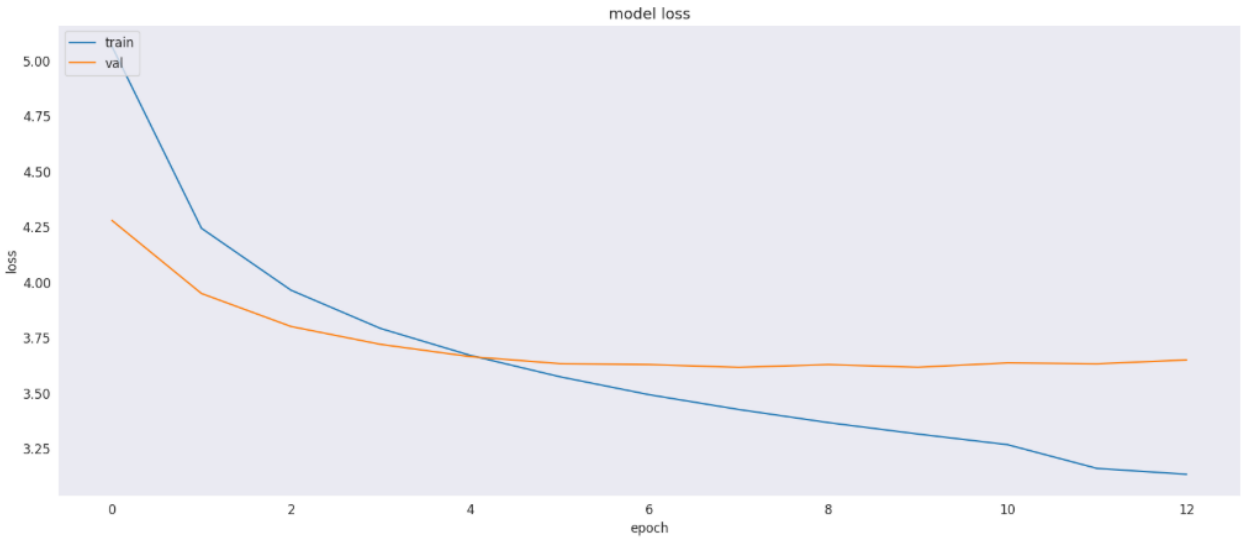

Figure 5: illustrates the loss rates of the model up to 12 epoch.

The trained model demonstrates the exemplary performance as compared to other caption generator based on DNNs such as ResNet, VGGNet, AlexNet.

5. Conclusion

The existing image caption generators have demonstrated the remarkable progress in generating grammatically correct captions but most of these are still suffering from the object hallucination. The benchmark scores such as BLEU, ROUGE are not adequate to check quality of the generated caption and measure their performance because they were built originally for machine translation. Although there was a debate on the qualitative measure on the context of generated caption, however some of the metric such as CIDEr and SPICE addresses the same. By taking into account of all these metrics we need to commensurate some other metrics to measure performance in hallucination detection. For this purpose there exist some novel metrics which includes CHAIR, FAITHSCORE and ALOHa.

In this work we have demonstrated a neuro-symbolic framework where the caption generator is successfully generating the desired caption as a sequence of words. These words are then mapped onto a vector called V2. Similarly, the proposed model also embeds a set of recognized objects from the input image in another vector V1. These two vectors are compared using the KMP pattern matching technique to detect repetition of object name or non-existence of any object or the object appeared in caption but not present in the vector V1.

Here we have combined the semantic representation of the generated caption with the deterministic matching capability of KMP algorithm. Unlike the traditional semantic analysis, which is purely rely on the probabilistic confidence factor, this KMP pattern matching provides a polynomial time complexity for substring matching which provides better adaptability in terms of hallucination detection. However, this work may be extended towards the incorporation of semantic knowledge graph along with synonyms matching for multilingual caption evaluation.

REFERENCES

1. Marc Tanti, Albert Gatt, and Kenneth P. Camilleri. Where to put the image in an image caption generator. Natural Language Engineering, 24:467–489, 2018.
2. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Ben-gio. Show, attend and tell: Neural image caption generation with visualattention. 32nd International Conference on Machine Learning, ICML 2015, 3:2048–2057, 2015.
3. Moses Soh. Learning CNN-LSTM architectures for image caption generation. Nips, pages 1–9, 2016.
4. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June:3156–3164, 2015.
5. Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, and Pushpak Bhattacharyya. A hindi image caption generation framework using deep learning. ACM Transactions on Asian and Low-Resource Language Information Processing, 20, 4 2021.
6. Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. 4 2024.
7. Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 10 2023.
8. Rongjie Li, Yu Wu, and Xuming He. Learning by correction: Efficient tuning task for zero-shot generative vision-language reasoning. 4 2024.
9. Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. 9 2023.
10. G. Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, and K. B.V.Dakshina Murthy. Domain-adaptive zero-shot image classification via clipand large language models. Journal of Engineering and Technology for Industrial Applications, 12:78–88,5 2026.
11. Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. 9 2024.
12. Donald E Knuth, James H Morris, and Vaughan R Pratt. Fast pattern matching in strings*. Technical report, SIAM Journal of Computing, Vol-6, No-2, June 1977.
13. Saswati Soumya Tripathy, Laxmipriya Barik, Prajnya Paramita Sahu, and Haraprasad Naik. Automatic threat detection using deep neural networks. IEEE, OCIT Proceeding. pages 66–71, 2022
14. H. Naik, P. Behera, S. Tripathy, and Laxmi Priya, "A Comprehensive Investigation on Image Caption Generation using Deep Neural Networks", *INFOCOMP Journal of Computer Science*, vol. 21, no. 1, Jun. 2022.
15. Lymperaiou, M., FIlandrianos, G., Dimitriou, A., Voulodimos, A., Stamou, G. (2026). HalCECE: A Framework for Explainable Hallucination Detection Through Conceptual Counterfactuals in Image Captioning. In: Guidotti, R., Schmid, U., Longo, L. (eds) Explainable Artificial Intelligence. xAI 2025.
16. Mengya Zhang, Yuan Zhang, and Qinghui Zhang. Attention-mechanism-based models for unconstrained face recognition with mask occlusion. Electronics (Switzerland), 12, 9 2023.

17. Roshni Padate, Amit Jain, Mukesh Kalla, and Arvind Sharma. Image caption generation using a dual attention mechanism. Engineering Applications of Artificial Intelligence, 123:106112, 2023.
18. Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. 3 2019
19. Raimonda Staniute and Dmitrij ˇSeˇsok. A systematic literature review on image captioning. Applied Sciences (Switzerland), 9, 2019.
20. Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob:4904–4912, 2017
21. Mario G´omez Mart´ınez Tutor, Anna Bosch Ru´e Profesor, and Jordi Casas Roma Valencia. Trabajo final de m ´Aster ´Area: Big data /machine learning deep learning for image captioning an encoder-decoder architecture with soft attention. Technical report, 2019
22. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:664–676, 2017.
23. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem:4565–4574, 2016.
24. Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena, 404:1–43, 2020.
25. Ling Zhao, Hanhan Deng, Linyao Qiu, Sumin Li, Zhixiang Hou, Hai Sun, and Yun Chen. Urban multi-source spatio-temporal data analysis aware knowledge graph embedding. Symmetry, 12:1–18, 2020