IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A MULTI MODEL CNN-LSTM FOR IMAGE CAPTION GENERATION

[1] Dr. A. Obulesh, [2]Sai Harshitha Aluru, [3]Tejaswi Sattarshetty, [4]Keerthi Yalamaddi, [5]Zaib Unnisa Nayeem

[1]HOD, [2]Student, [3]Student, [4]Student, [5]Student

[1]Artificial Intelligence,
[1]Vidya Jyothi Institute Of Technology, Hyderabad, India

Abstract: Automatically generating descriptive captions for images is indeed a crucial task for machines to understand and describe natural scenes accurately. Current state-of-the-art models utilize deep convolutional neural networks (CNNs) to extract visual features from the images, followed by recurrent neural networks for the task of image caption generation. However, these models often overlook the textual information present in images, limiting the effectiveness of the generated captions. This study focuses on utilizing the text present in the images along with the visual features extracted from CNN. In this paper, we propose integration of deep convolutional neural networks with long-short term memory to generate context related captions. We achieve this by fusing the textual features obtained from text along with the visual features obtained from the CNN model. We use the benchmark Flickr8k dataset for validating the efficiency of proposed approach. The experiments reveal that fusing of visual-textual features performs better than existing state-of-the-art models.

IndexTerms **– Image captioning, Convolutional neural networks, Long short term memory, Visual cues, Textual cues**

I. Introduction

Human's naked eye is capable of describing the natural scenes with ease and effectiveness in one glance. Our visual processing effortlessly translates a multitude of elements – objects, colors, their spatial relationships, and interactions – into a coherent narrative. Unlike humans, machines tend to face a significant challenge while describing the natural scenes effectively. Image captioning emerges as a bridge between these two worlds, aiming to automatically generate textual descriptions that not only bound to grammatical rules but also capture the essence of the scene, mirroring human perception. One of the primary challenges lies in the intricacies of understanding natural scenes. Unlike structured data or text, images contain a vast array of visual elements, including objects, colors, textures, spatial relationships, and more. Machines must first identify and categorize these elements within the image, which involves complex processes such as object detection, semantic segmentation, and feature extraction.

However, mere identification of objects and their attributes is not sufficient for effective image captioning. Machines must also grasp the overall context of the scene and the relationships between different elements. This includes understanding actions, interactions, scenes, and even abstract concepts that may be depicted in the image. For instance, recognizing that a person is "riding a bicycle" involves not just identifying the person and the bicycle but also understanding the dynamic relationship between them. Translating this understanding into natural language descriptions adds another layer of complexity. Machines must generate grammatically correct sentences that not only describe the visual content accurately but also convey the semantics and essence of the scene in a coherent and human-like manner. This involves not only linguistic knowledge but also an understanding of context, pragmatics, and common-sense reasoning. Further, they also have to translate this understanding into meaningful sentences by giving natural descriptions, making the task more complex. Some potential applications for automatic image captioning include having natural conversations with devices [1-3], instantly understanding medical scans [4-6], improving traffic management through descriptions [7], and even bringing visual experiences to the blind [8-12].

For the task of caption generation the recent state-of-the-art methods use deep neural network for feature extraction followed by an encoder-decoder pipeline for machine translation [13-17]. They provide better captions as compare to traditional approaches. Flickr8k [18] is a benchmark dataset for image captioning. However, in all of the studies explored till today, they capture the visual features only, leaving some semantic information like salient text present in the images like building names, number plate, street signs etc. This text can provide important information which can be used for generating captions for images [19].

In this study we elevate the image captioning process by integrating the textual data embedded within the image itself. While recent deep neural network (DNN) based approaches have achieved some success in image captioning, they largely depend on visual features only. We believe that some natural scenes containing semantic information like signs, license plates, and building names are conveyed only through textual elements present in those images. By incorporating these "textual features" along with the visual features, we propose an approach to achieve fine-grained image captioning, resulting in descriptions that align more closely with human perception and understanding.

In this paper we first discuss about the literature survey in section IIII, followed by the proposed research methodology in section III. We then talk about the results obtained in section IV, and conclude in the section V.

II. LITERATURE SURVEY.

Image captioning has been a challenging task in computer vision. Before using neural networks, it was problematic to generate image captions using two conventional methods: first is retrieval-based and second is template based. In retrieval based-methods [20-22] the system searches through captions and selects the most suitable image. In template-based methods [23-25], the system utilizes template model wherein image is analysed and relevant information is added to model to generate caption. Though not ideal, these methods paved the way for advanced techniques being used today. Recent breakthroughs in deep learning [26] have made captioning much easier. In a study by Socher et al. [27] authors proposed a new method that uses two deep learning models. The first model analyses the image input whereas the other model relies on understanding captions using dependency tree structure. After this the system searches for a match. Finally based on commonality between image and captions, system generates a new caption. Karpathy et al. [28] recommended sentence numbering system. Here, sentences and images are implanted into similar area.

Convolutional neural network(CNN) method was implemented in [29] consisting of tree relations and their dependencies. The caption and image is converted to vector depicting key features. Using this vector, we place image and caption together allowing to compare the similarity in image and caption in depth and choosing ideal caption. Ma et al. introduced a method using multimodal Convolutional neural network [30]. In this, the network simultaneously analyses images and their captions. Initially, image is briefly scanned to identify its characteristics followed by sentence analysis. Then the network compares them at different levels of details and allocates them a score. This score reflects how well caption matches the image. Ultimately using this information, captions are generated more accurately. Kiros et al. [31] introduced a language model where multimodal methods were replaced by log-bilinear language approach. natural language processing is used to generate a word based on words produced in the previous step. In research paper by Mao et al. [32], a sequence model recurrent neural networks (RNN) was implemented in order to create original captions. Another paper [33] used deep Convolutional neural network to extract visual cues from image followed by recurrent neural networks model to create captions. Long short-term memory (LSTM) along with deep Convolutional neural network was implemented by Vinyals et al. [34] that generated captions using encoder decoder functionality.

III. RESEARCH METHODOLOGY

This research paper presents a methodology for image captioning by utilizing the textual content present in the images along with the visual features of the image. We give the details of dataset used followed by giving an overview of CNN and LSTM. We then describe the proposed model for caption generation.

3.1 Dataset

Flickr8k [18] dataset is used for the experiments. In this dataset, there are five annotated sentences for each and every image. This is extracted from Flickr with a total of 8000 images. For the purpose of experiments, a standard split is used where 6000 images are used for training, 1000 images are used for validation, and 1000 images are used for testing. Notably, Flickr8k's diversity makes it ideal for our task as it contains images with both visual and textual content (approximately 1,722) alongside purely visual ones (around 6,369). Details on this distribution can be found in Table 1.

Table 1: Distribution of Flickr8k dataset.

Total images	Size			Visual + textual Cue images	Images with only textual cue
	Train	Validation	Test		
8091	6091	1000	1000	1722	6369

3.2 Convolutional Neural Networks

Deep Convolutional neural networks (CNNs) [35] are specifically designed for understanding images and videos. The dense architecture of CNN learns to recognize patterns and features in images. In our proposed model, instead of training a deep convolutional neural network from scratch, we utilize a pre-trained VGG Net [36] model to extract the visual representations.

3.3 Long Short-Term Memory

Long Short-Term Memory (LSTM) [37] are used for the task of sequence generation [38]. Its architecture consists of three gates: input gate (ig), output gate (og), forget gate (fg). The input (ig) and output (og) gates in LSTM are used to control the input and output respectively. The forget gate (fg) decides whether or not to forget the current cell value.

The mathematical functions of these gates and memory cell are given below (3.1-3.6):

$$ig_t = \sigma(W_{igx}x_t + W_{igm}m_{t-1}) \quad (3.1)$$
$$fg_t = \sigma(W_{fgx}x_t + W_{fgm}m_{t-1}) \quad (3.2)$$
$$og_t = \sigma(W_{ogx}x_t + W_{ogm}m_{t-1}) \quad (3.3)$$
$$cg_t = fg_t \otimes cg_{t-1} + ig_t \otimes h(W_{cgx}x_t + W_{cgm}m_{t-1}) \quad (3.4)$$
$$m_t = og_t \otimes cg_t \quad (3.5)$$
$$P_{t+1} = Softmax(m_t) \quad (3.6)$$

where σ(.) and h(.) denotes sigmoid activation function and hyperbolic tangent activation function respectively. W denotes matrices of trained parameters, and ⊗ denotes gate value product. In order to get the probability P_t we use SoftMax activation function in the last layer of LSTM.

3.4 Proposed Model for Image Caption Generation

Recent advancements in image captioning have primarily focused on visual features alone which give the generic description of the scene in the image. But they lack in producing specific details of the scene. In our proposed model, we are integrating both visual image features and textual features for fine-grained image caption generation.
Figure 1 represents the overview of the proposed model. This consists of three modules. The initial module of the model utilizes the Visual Geometry Group (VGG) net, a deep convolutional neural network (CNN) model, to extract visual features (referred to as visual cues) from the images. In the second module, KerasOCR [39], an Optical Text Recognition (OCR) method is used to extract the salient text from the image. Further, we use a word embedding algorithm, Word2Vec [40] to represent text features in vector form (referred to as textual cues). This algorithm maps the words into a continuous vector space, where semantically similar words are inserted adjacent to one another. In the third module, we use the fusion of visual and textual cues in order to generate captions through LSTM. This results in more informative and contextually relevant image captions.

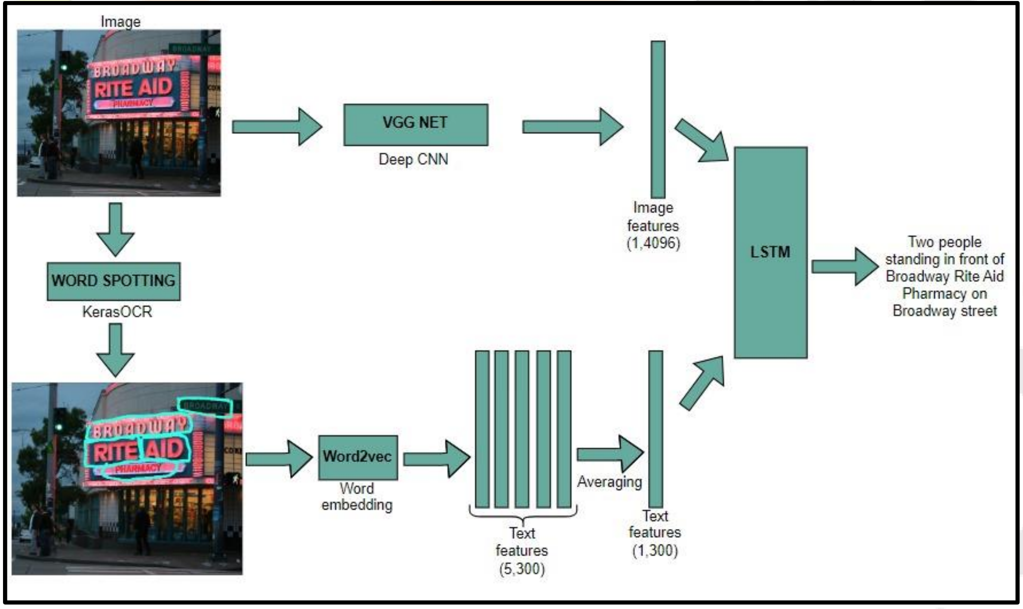

Figure 1: Proposed model for image captioning in a natural scene.

A) Module-1: Extraction of Visual cues

In the proposed framework, we adopt the VGG16 a pre-trained model, which is trained on millions of images from the ImageNet [41] dataset and further fine tune it. This architecture consists of 3x3 convolutional layers stacked in increasing depth, augmented by max-pooling for down sampling. The network culminates into two fully connected layers with 4096 nodes each, followed by a SoftMax classifier responsible for classifying images into 1000 classes. To obtain the visual cues [40], a 4096-dimensional feature vector is extracted from the second last layer of the network. By this we obtain a (1,4096) dimensional visual cue vector for each image I_m.

B) Module-2: Extraction of Textual cues

In this module we first use the KerasOCR library to extract the text present in the image I_m. The extracted text is converted to vectors with the help of Word2Vec algorithm. Each spotted word t_i in an image is represented using a word vector ft_i of 300 dimensions. If there are N spotted words, the textual cue size is 300 x N. In order to maintain the consistency of dimensions of textual cues across the images we average the N features. By this we obtain a (1,300) dimensional textual cue vector for each I$_m$.

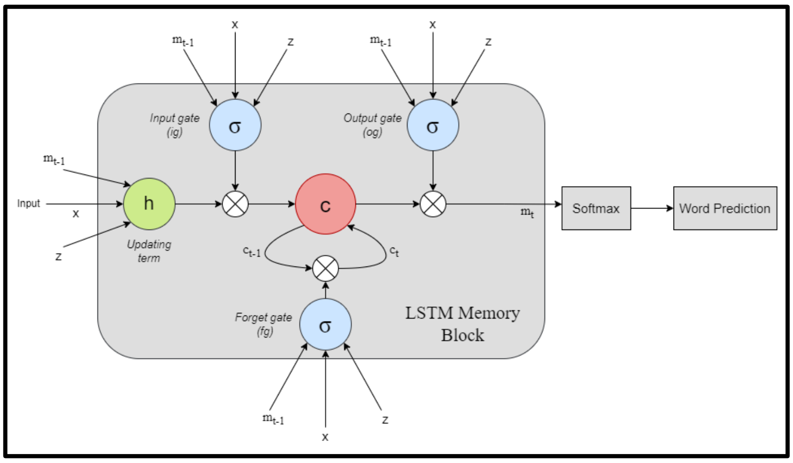

Figure 2: Long short-term memory block diagram

C) Module-3: LSTM based caption generator

Feature fusion: The proposed method introduces a multimodal fusion approach that combines both visual and textual cues. The visual cues of dimension (1, 4096) obtained from module 1 are denoted as ‘x’. The textual cues of dimension (1,300) obtained from module 2 are denoted as ‘z’. In the fusion process, the extracted visual and textual feature vectors are together fed into the LSTM model to generate image captions.

LSTM: As described in section 3.2, the inputs to the LSTM gates consists of current visual cue ‘x’ and the previous hidden state ‘m$_{t-1}$’, representing the contextual information from the previous time step. A context vector ‘z’ is introduced which represents the textual cues extracted from the input image. As shown in Figure 2, by integrating ‘z’ as an additional input to the LSTM gates, the semantic information provided by z enriches the image information available to the LSTM model during caption generation. This enables the model to consider not only the visual features of the image but also the semantic context provided by the textual cues. Adam[43] is used as optimizer.

The proposed model’s mathematical formulation of the gates and memory cell is given below (3.7 – 3.12):

$$ig_t = \sigma\left(W_{igx}x_t + W_{igm}m_{t-1} + W_{igz}Z\right) \tag{3.7}$$

$$fg_t = \sigma\left(W_{fgx}x_t + W_{fgm}m_{t-1} + W_{fgz}Z\right) \tag{3.8}$$

$$og_t = \sigma\left(W_{ogx}x_t + W_{ogm}m_{t-1} + W_{ogz}Z\right) \tag{3.9}$$

$$cg_t = fg_t \otimes cg_{t-1} + ig_t \otimes h\left(W_{cgx}x_t + W_{cgm}m_{t-1} + W_{cgz}Z\right) \tag{3.10}$$

$$m_t = og_t \otimes cg_t \tag{3.11}$$

$$P_{t+1} = Softmax(m_t) \tag{3.12}$$

where z is the context vectors that represent the text features, and other parameters are same as described in section 3.2. Unlike as mentioned in section 3.2, the common parameters for LSTM are x and z for each time step t. The start word and stop word of the sentence are identified by the two special terms, S$_0$ and S$_N$. To signify that the generated caption is successful, LSTM employs the stop word.

IV. RESULTS AND DISCUSSION

We perform experiments on the Flicker8k dataset and analyze the performance of proposed (visual + textual) model by comparing it with the baseline (visual) model with the help of BLEU [44] scores and captions generated as shown in Table 2, Figure 3 and Figure 4 respectively. Table 2 explains the BLEU (1-4) scores that are obtained from both the models i.e. baseline model and proposed model. In Figure 3, we compare the image captions generated by both the models when there is no text present in images. In Figure 4, we compare the image captions generated by the existing model and suggested model where some text is present in the images. We first gave an overview of evaluation metrics used, followed by comparison of BLEU scores. Finally, we delve into comparing the image captions generated by both the models.

4.1 Evaluation Metrics:

BLEU (Bilingual Evaluation Understudy) score uses n-gram based metrics to assess the quality of generated captions. For instance, BLEU calculates precision scores for n-grams of different lengths, ranging from single word called BLEU-1 (unigram) and four words referred to as BLE-4(four-gram). Through that each BLEU score represents how much similarity exists between a candidate and reference caption in terms of n-gram overlap. In evaluation, use BLUE 1 TO BLUE 4.

4.2 Comparison of BLEU scores generated by proposed model with baseline model:

Table 2 depicts the BLEU scores with the baseline and proposed models. When only visual cues are used, we obtain 50.1, 30.47, 19.49 and 11.65 as the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores respectively. But after integration of textual cues with visual cues, the BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores are found to be 56.5, 34.35, 22.02, and 13.5 respectively. The proposed model showed an improvement of 6.4, 3.88, 2.53 and 1.85 in BLEU (1-4) scores respectively when compared to baseline model. This clearly indicates that, incorporating textual features while caption generation can generate captions with a higher quality.

Table 2: BLEU scores of baseline model and proposed model

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline model (only visual cues)	50.1	30.47	19.49	11.65
Proposed model (Visual and Textual Cues)	56.5	34.35	22.02	13.5

4.3 Comparison of captions generated by proposed model with baseline model:

In order to maintain generalizability, we show a comparison of the captions generated with both baseline and proposed model in two scenarios. One, where the images do not containing any text and two where the images contain text within it. In both the cases both textual and visual cues are used for training but if there is no text present, we consider a zero vector as textual cue.

a) Captions for images without text:

Figure 3 displays the captions generated for images without text. Figure 3(a) shows the captions generated by the baseline model and Figure 3(b) shows the captions generated by our proposed model. Even though these images do not contain any text, the proposed model is able to generate better captions because it is integrated with extra information (textual cues) which conveys the presence of text within the image or not. For example in Figure 3(a), “two children are playing in the tent” is generated by baseline and “little girl in pigtails sitting in front of rainbow painted rainbow” is given by the proposed one. The second caption gives accurate information of what is happening in the image.

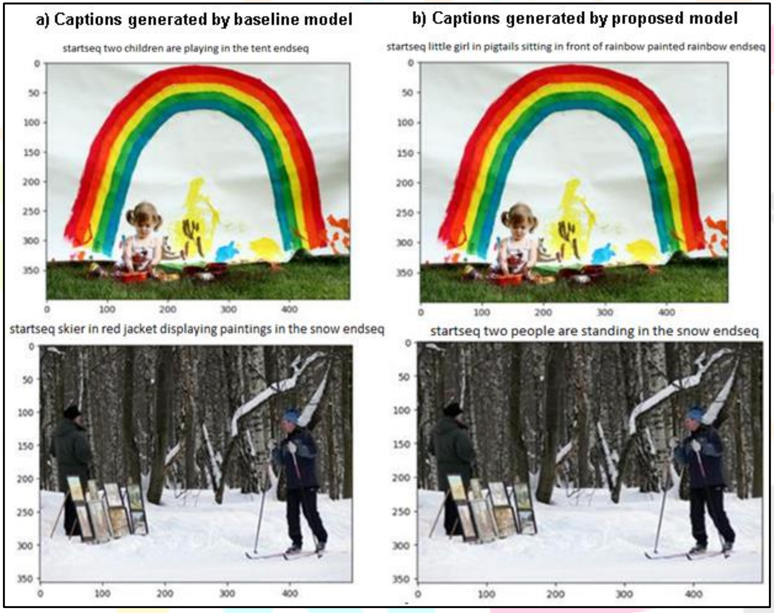
a) Captions generated by baseline model
b) Captions generated by proposed model
startseq two children are playing in the tent endseq
startseq little girl in pigtails sitting in front of rainbow painted rainbow endseq
startseq skier in red jacket displaying paintings in the snow endseq
startseq two people are standing in the snow endseq
0
50
100
150
200
250
300
350
0
100
200
300
400

Figure 3: Generated captions for images without inbuilt text.

b) Captions for images with text:

Figure 4 contains the captions predicted for the images with text. Figure 4(a) shows image captions generated by baseline model and Figure 4(b) shows image captions generated by proposed model. The caption generated by baseline was “man in red jacket is standing in front of the street”. But there is no man with red jacket. Instead it considered the red colored board in the background as the color of the jacket. The resulting caption was found to be “two people standing in front of broadway rite aid pharmacy on broadway street” with the proposed model.

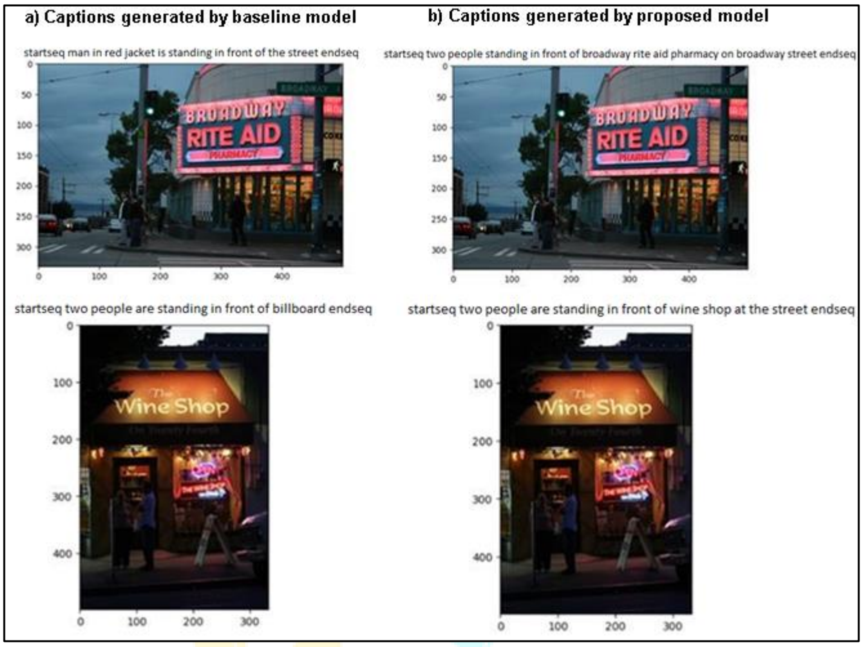
a) Captions generated by baseline model
b) Captions generated by proposed model
startseq man in red jacket is standing in front of the street endseq
startseq two people standing in front of broadway rite aid pharmacy on broadway street endseq
startseq two people are standing in front of billboard endseq
startseq two people are standing in front of wine shop at the street endseq
0
50
100
150
200
250
300
400

Figure 4: Generated captions for images with inbuilt text.

The effectiveness of incorporating textual cues into the model is proven through Table 2, Figure 3 and Figure 4. The proposed model which uses both textual and visual cues into consideration demonstrates deeper understanding of the natural scene than the baseline model which considers only visual cues. BLEU scores also show a similar trend. We believe that, identifying text and incorporating it with LSTM helps the deep learning model to understand the natural scene more effectively.

V. CONCLUSION

We propose a novel image captioning model that employs both visual and textual information present in images. We propose that if the model is integrated with just visual features it lacks the ability to capture whole semantics of a scene. Hence, our technique including the text data (salient text) found in images to the model. The proposed method was tested on benchmark Flickr8k dataset. When compared with other state-of-the-art models which use only visual information, the proposed approach gave better results. For future work we aim to generate task-specific captions.

REFERENCES

[1] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer International Publishing, 2020.

[2] Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." arXiv preprint arXiv:1606.01847 (2016).

[3] Zhang, Pengchuan, et al. "Vinvl: Revisiting visual representations in vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[4] Pavlopoulos, John, Vasiliki Kougia, and Ion Androutsopoulos. "A survey on biomedical image captioning." *Proceedings of the second workshop on shortcomings in vision and language*. 2019.

[5] Huang, Jia-Hong, Ting-Wei Wu, and Marcel Worring. "Contextualized keyword representations for multi-modal retinal image captioning." *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 2021.

[6] Ayesha, Hareem, et al. "Automatic medical image interpretation: State of the art and future directions." *Pattern Recognition* 114 (2021): 107856.

[7] Li, Wei, et al. "The traffic scene understanding and prediction based on image captioning." *IEEE Access* 9 (2020): 1420-1427.

[8] Gurari, Danna, et al. "Captioning images taken by people who are blind." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer International Publishing, 2020.

[9] Sidorov, Oleksii, et al. "Textcaps: a dataset for image captioning with reading comprehension." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer International Publishing, 2020.

[10] Makav, Burak, and Volkan Kılıç. "A new image captioning approach for visually impaired people." *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2019.

[11] Dognin, Pierre, et al. "Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge." *Journal of Artificial Intelligence Research* 73 (2022): 437-459.

[12] Ahsan, Hiba, et al. "Multi-modal image captioning for the visually impaired." *arXiv preprint arXiv:2105.08106* (2021).

[13] Momtaz, Mozhgan, et al. "Multi-modal deep fusion for bridge condition assessment." *Journal of Infrastructure Intelligence and Resilience* 2.4 (2023): 100061.

[14] Chen, Jin, et al. "When large language models meet personalization: Perspectives of challenges and opportunities." *arXiv preprint arXiv:2307.16376* (2023).

[15] Rodrigues, Douglas, et al. "Fine-Tuning Restricted Boltzmann Machines Using No-Boundary Jellyfish." *VISIGRAPP (4: VISAPP)*. 2023.

[16] Bai, Shuang. "Growing random forest on deep convolutional neural networks for scene categorization." *Expert systems with applications* 71 (2017): 279-287.

[17] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

[18] Rashtchian, Cyrus, et al. "Collecting image annotations using amazon's mechanical turk." *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. 2010.

[19] Gupta, Neeraj, and Anand Singh Jalal. "A comparison of visual attention models for the salient text content detection in natural scene." *2017 conference on information and communication technology (CICT)*. IEEE, 2017.

[20] Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer Berlin Heidelberg, 2010.

[21] Ordonez, Vicente, Girish Kulkarni, and Tamara Berg. "Im2text: Describing images using 1 million captioned photographs." *Advances in neural information processing systems* 24 (2011).

[22] Gupta, Ankush, Yashaswi Verma, and C. Jawahar. "Choosing linguistics over vision to describe images." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 26. No. 1. 2012.

[23] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." *IEEE transactions on pattern analysis and machine intelligence* 35.12 (2013): 2891-2903.

[24] Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012.

[25] Ushiku, Yoshitaka, et al. "Common subspace for model and similarity: Phrase learning for caption generation from images." *Proceedings of the IEEE international conference on computer vision*. 2015.

[26] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[27] Socher, Richard, et al. "Grounded compositional semantics for finding and describing images with sentences." *Transactions of the Association for Computational Linguistics* 2 (2014): 207-218.

[28] Karpathy, Andrej, Armand Joulin, and Li F. Fei-Fei. "Deep fragment embeddings for bidirectional image sentence mapping." *Advances in neural information processing systems* 27 (2014).

[29] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[30] Ma, Lin, et al. "Multimodal convolutional neural networks for matching image and sentence." *Proceedings of the IEEE international conference on computer vision*. 2015.

[31] Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models." *International conference on machine learning*. PMLR, 2014.

[32] Mao, Junhua, et al. "Explain images with multimodal recurrent neural networks." *arXiv preprint arXiv:1410.1090* (2014).

[33] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60.6 (2017): 84-90.

[34] Verma, Akash, et al. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83.2 (2024): 5309-5325.

[35] Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." Biological cybernetics 36.4 (1980): 193-202.

[36] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[37] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[38] Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013).

[39] Poudel, Utsav, et al. "Applicability of OCR Engines for Text Recognition in Vehicle Number Plates, Receipts and Handwriting." *Journal of Circuits, Systems and Computers* 32.18 (2023): 2350321.

[40] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013)..

[41] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115 (2015): 211-252.

[42] Sharif Razavian, Ali, et al. "CNN features off-the-shelf: an astounding baseline for recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014

[43] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[44] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.