Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)

VOLUME:8, ISSUE:1, FEBRUARY 2025, PAPER NUMBER:40

LSTM BASED TENSOR-FLOW MECHANISM FOR GENERATING CAPTIONS FOR IMAGES

K Manohar
dept Computer *Science and Engineering*
Institute of Aeronautical Engineering
Hyderabad, India
20951a0584@iare.ac.in

K Nehath Varma
dept Computer Science and engineeing
Institute of Aeronautical Engineering
Hyderabad, India
20951a05a0@iare.ac.in

L Satwik
dept Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad, India
21955a0517@iare.ac.in

Dr. R Obulakonda Reddy
dept Computer *Science and Engineering Cyber Security*

Designation Professor
Institute of Aeronautical Engineering
Hyderabad, India
r.obulakondareddy@iare.ac.in

Abstract— This research explores the intersection of computer-vision and natural language processing(NLP) specifically focusing on image-captioning by combining convolutional neural networks and long short term memory models we aim to reveal the narratives embedded in visual content using xception a pre-trained CNN we map out elements such as objects textures and spatial relationships within an image these features serve as input for the LSTM which is a RNN architecture to generate descriptive stories considering the visual representation our goal extends beyond mere object identification instead we aspire to capture an images essence including its unspoken emotions and nuanced context to achieve this objective out model is trained on carefully curated image captioning datasets which provide valuable lessons in visual storytelling techniques.

Keywords--: Image-captioning, Convolutional-Neural Networks, Long Short Term Memory, Recurrent-Neural Network

I. INTRODUCTION

In the burgeoning realm of artificial intelligence, where machine perception stives to emulate human understanding, image captioning emerges as a captivating intersection of computer vision and natural language processing. This research delves beyond mere object recognition; it aspires to unveil the narratives woven within visual tapestries, empowering images to whisper their stories into the receptive ears of technology.

This quest holds profound significance not only for accessibility, where whispered descriptions illuminate the world for the visually impaired, but also for the broader landscape of human-computer interaction. Imagine traversing vast image repositories with the evocative eloquence of natural language queries, unearthing the perfect recipe photo by describing its culinary essence, or reliving cherished memories through automatically captioned personal photographs. These are but a glimpse of the transformative possibilities awaiting us in this burgeoning domain.

Fueled by the potent synergy between Long Short-Term Memory (LSTM) networks and the TensorFlow framework, our research embarks on a journey to unlock the hidden language of images. This pursuit transcends the confines of academic exploration; it represents a quest to bridge the gap between human and machine perception, allowing technology to see the world as a collection of pixels follow.

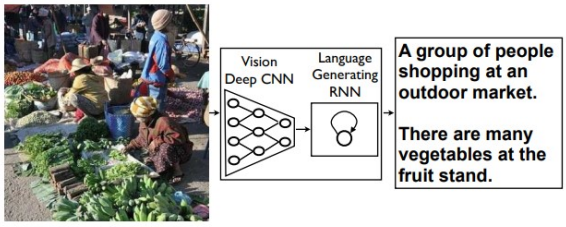


Figure 1. Our model, operates entirely on a neural network architecture comprising a vision CNN followed by a language-generating RNN. It translates input images into coherent natural language sentences, as illustrated in the example provided above.

In previous approaches to generating image descriptions, the common strategy involved piecing together various components, such as identifying groups of people at a market or recognizing specific items like vegetables at a fruit stand. These attempts typically involved combining solutions for different sub-problems. However, in contrast to this fragmented approach, our work introduces a unified model capable of processing images and generating descriptions seamlessly. This model, termed the Neural-Image Caption (NIC), operates in a way that, taking an image as input and learning to produce descriptive sentences that accurately depict the content of the image.

Our inspiration stems from recent advancements in machine translation, particularly the shift towards simpler yet highly effective approaches using Recurrent Neural Networks (RNNs). Traditionally, machine translation involved multiple sequential tasks, such as translating individual words, aligning them, and reordering them. However, modern techniques employing RNNs have demonstrated remarkable performance by condensing the translation process into a more streamlined framework. In this configuration, an "encoder" RNN analyzes the source sentence, transforming it into a concise fixed-length vector representation. This vector acts as the initial reference for a "decoder" RNN, which then produces the target sentence in the specified language.

Following this elegant paradigm, we propose substituting the encoder RNN with a deep Convolutional-Neural Network in our image captioning model. CNNs have proven their capability in extracting rich representations from input images, typically pre-training is conducted for classification of images. By leveraging the last hidden layer of the CNN serving as the input to the RNN decoder, our model, NIC, can effectively generate descriptive sentences (see Fig. 1) derived from the visual information within the input image.

Utilizing a CNN as an "image encoder" offers a natural synergy between perception and language tasks. By embedding the image into a representation of a fixed length vector, the CNN facilitates various vision-related objectives, including image classification and, in our case, caption generation. This unified approach streamlines the process of image description generation, eliminating the need for disjointed sub-components and reaching the forefront of performance in the task. Thus, NIC represents a novel and effective solution for generating natural language descriptions from images.

II. Related work

Within the field of computer vision, the task of producing descriptive natural language from visual information presents a challenge, historically focused more on video content rather than still images. Early attempts at addressing this task often resulted in complex systems comprising primitive visual recognizers paired with formalized structured languages, like And-Or Graphs or logical systems. These systems, however, relied heavily on manual design and were relatively rigid, typically demonstrated only in specific domains like traffic scenes or sports. Consequently, while effective within their limited scope, they lacked the flexibility and generalizability required for broader applications.

More recently, interest has surged in describing still images using natural language, leveraging advancements in object recognition techniques. Approaches have emerged that utilize detections utilizing information about objects, their attributes, and spatial relationships to propel natural language generation systems. For instance, some methods infer scene elements from detections and convert them into text using predefined templates. Others piece together descriptions by incorporating detected objects and relationships extracted from images. Despite these advancements, many existing approaches still suffer from limitations in expressivity and are heavily reliant on handcrafted design, hindering their ability to generate diverse and nuanced descriptions.

A significant body of research has also focused on ranking descriptions for given images, founded on the idea of embedding both images and text within a unified vector space. These approaches aim to fetch descriptions closely aligned with a specified image query in the embedding space. While neural networks have been employed for co-embedding images and sentences, they often fall short in generating novel descriptions, particularly for previously unseen compositions of objects.

In this context, our work proposes a novel approach that merges deep convolutional neural networks for image classification with recurrent networks for sequence modeling. This unified network architecture enables the generation of descriptions for images in a holistic manner, drawing inspiration from the achievements of sequence generation in machine translation. Unlike previous methods that either use feedforward neural networks or provide text input directly without visual context, our model leverages a recurrent neural network with enhanced capabilities and integrates visual input directly into the sequence generation process. This enables the network to proficiently monitor the objects referenced in the generated text, leading to substantially improved results on established benchmarks compared to prior works.

While related efforts have explored similar directions, our approach distinguishes itself by its more sophisticated RNN model and the direct integration of visual input into the sequence generation process. Additionally, unlike approaches that construct joint multimodal embedding spaces with separate pathways for images and text, our model operates as a single network, facilitating more coherent and contextually aware descriptions.

III. Proposed Model

In the context of our research, we introduce a neural and probabilistic framework for generating descriptions directly from images, drawing inspiration from recent advancements in statistical machine translation. Leveraging the effectiveness of end-to-end optimization in achieving Cutting-edge outcomes, we adopt a similar approach using powerful sequence models, particularly recurrent neural networks (RNNs). These models encode variable-length input sequences into fixed-dimensional vectors and decode them into desired output sentences. Extending this principle to image description generation, we input images into the network and employ the same concept of "translating" them into descriptive sentences. By encoding visual content into fixed-dimensional vectors, our framework seamlessly integrates both training and inference processes, facilitating improved performance and more accurate descriptions.

Therefore, we suggest maximizing the probability of the accurate description directly from the image by employing the following formulation:

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta) \qquad (1)$$

Here, θ denotes the parameters of our model, I represents an image, and S denotes its correct transcription. As S can encompass any sentence, its length is unrestricted. Consequently, it's typical to employ the chain rule to depict the joint probability across S0, . . . , SN , where N is the length of this particular example as

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}) \qquad (2)$$

Here, we've omitted the dependence on θ for simplicity. During training, (S, I) constitutes a pair of training examples, and we iteratively enhance the sum of log probabilities outlined in (2) across the entire training dataset using stochastic gradient descent.

Utilizing a Recurrent Neural Network (RNN) to model p(St|I, S0, . . . , St−1) is common practice, wherein the dynamic number of words conditioned upon up to t − 1 is represented by a fixed-length hidden state or memory ht. This memory undergoes updates upon encountering a new input xt through the application of a non-linear function f:

$$h_{t+1} = f(h_t + x_t) \qquad (3)$$

In order to enhance the clarity of the aforementioned RNN, two significant design decisions need to be addressed: determining the precise structure of f and establishing the methodology for inputting images and words as xt. For f, we opt for a Long-Short Term Memory (LSTM) network, renowned for its exceptional performance in sequential tasks like translation. Further elaboration on this model is provided in the subsequent section.

we utilize a Convolutional Neural Network (CNN) to represent images, a widely utilized and researched architecture known for its excellence in recognition and detection of objects, with our chosen CNN utilizing innovative batch normalization techniques and achieving top performance in the ILSVRC 2014 classification competition. CNNs also exhibit transfer learning capabilities, extending their effectiveness to tasks like scene classification. Meanwhile, words are depicted utilizing an embedding model.

IV. Sentence Generator Based on LSTM

The choice of function f in equation (3) relies on its ability to address issues such as vanishing and exploding gradients, which are common hurdles encountered in the design and training of RNNs. To tackle these obstacles, a particular kind of recurrent neural network, referred to as LSTM, was introduced and has demonstrated its effectiveness, highly successful in tasks such as translation and sequence generation. At the heart of the LSTM model is a memory cell denoted as c, which stores information at every time step about the inputs observed until that moment.

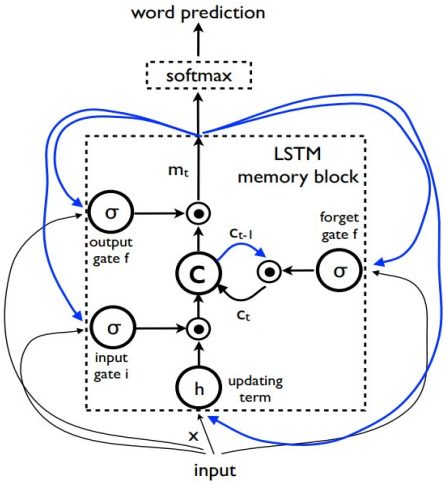


Figure 2.

The behaviour of this cell is governed by "gates," which are layers that function multiplicatively and decide whether to preserve or discard a value depending on the gate's state. Specifically, three gates are utilized: the forget gate f determines whether to forget the current cell value, the input gate 1 governs the integration of new information, and the output gate 0 controls the output of the cell. This structure adeptly tackles the issues of vanishing and exploding gradients encountered in RNNs.

Figure 2: LSTM architecture involves a memory block that encompasses a cell, denoted as c, regulated by three gates. In the provided visualization, the recurrent connections, depicted in blue, illustrate the feedback mechanism where the output m at time t - 1 is looped back to the memory at time t through the three gates. Specifically, the cell value is fed back via the forget gate, while the predicted word at time t - 1 is additionally fed back, along with the memory output m at time t, into the Softmax for word prediction.

These gates determine whether to read the input (input gate i) and whether to output the new cell value (output gate o). The gates' definitions, along with the cell update and output, are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1}) \quad (4)$$
$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \quad (5)$$
$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1}) \quad (6)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \quad (7)$$
$$m_t = o_t \odot c_t \quad (8)$$
$$p_{t+1} = \text{Softmax}(m_t) \quad (9)$$

In this context, ⊙ symbolizes the product operation with a gate value, while the different W matrices serve as parameters that undergo training. The use of multiplicative gates, alongside sigmoid σ (•) and hyperbolic tangent h(•) nonlinearities, facilitates robust training of the LSTM by effectively addressing issues related to exploding and vanishing gradients. The resulting equation mt is then fed into a Softmax function to generate a probability distribution pt across all words.

Training the Model:

The training of the LSTM model involves predicting each word of the sentence subsequent to processing the image and all preceding words, adhering to the conditional probability distribution p(St|I, S0, . . . , St−1). Conceptually, we can envision the LSTM in its unrolled representation, where a succession of LSTM memory cells is sequentially generated to handle each word in the sentence in conjunction with the image and preceding words.

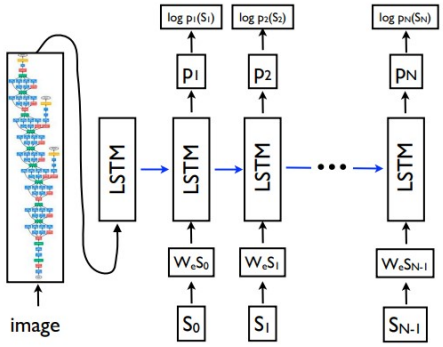


Figure 3. The LSTM model is integrated with a CNN image embedder and word embeddings. The unrolled connections between LSTM memories are depicted in blue, aligning with the recurrent connections illustrated in Figure 2. All LSTMs utilize identical parameters.

In the unrolled version, each LSTM processes the input image I and each word in the sentence S = (S_0, . . . , S_N) sequentially, all LSTMs utilize identical parameters. The output mt−1 from the LSTM at time t − 1 serves as input for the LSTM at time t, facilitating information flow between consecutive time steps (as depicted in Figure 3). Throughout this unrolling process, all recurrent connections transform feed-forward connections. To specify, if we designate the input image as S = (S_0, . . . , S_N), the unrolling procedure progresses in a manner where each LSTM unit processes both the image and every word in the sentence, ensuring continuity and coherence in the generated descriptions.

$$x_{-1} = \text{CNN}(I) \quad (10)$$
$$x_t = W_e S_t, \quad t \in \{0 \ldots N-1\} \quad (11)$$
$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \ldots N-1\} \quad (12)$$

In our framework, Each word is represented as a one-hot vector St, with dimensions corresponding to the size of the dictionary. Specifically, S0 is designated as a special start word and SN as a special stop word, denoting the beginning and end of the sentence respectively. When the stop word is emitted, it signals to the LSTM that a complete sentence has been generated. Both the image and the words are mapped into the same space, accomplished by utilizing a vision CNN for the image and word embedding We for the words. The image I is input only once at t = -1 to inform the LSTM about the image contents. Empirical testing has revealed that introducing the image at each time step as an additional input results in subpar performance, as it permits the network to potentially exploit noise in the image, thereby leading to overfitting.

V. Results

A. Datasets Desciption

For evaluation purposes, we employ several datasets comprising images paired with English descriptions. With the exception of the SBU dataset, each image within these datasets is accompanied by five sentences intended to offer relatively impartial and visually descriptive annotations. However, the SBU dataset stands out due to its reliance on descriptions provided by image owners upon upload to Flickr, which lack assurances of visual accuracy or impartiality, resulting in a higher level of noise.

The statistics of the datasets are as follows:

Typically, Pascal dataset is designated for testing purposes exclusively following the training of a system on various datasets, including any of the other four datasets referenced. This applies similarly to the SBU dataset, we allocate 1000 images for testing purposes and utilize the remaining data for training. following a similar approach as previous studies. Similarly, for the MSCOCO validation set, we reserve 4K random images for testing, referred to as COCO-4k, which we utilize to present results in the subsequent sections.

B. Evaluation

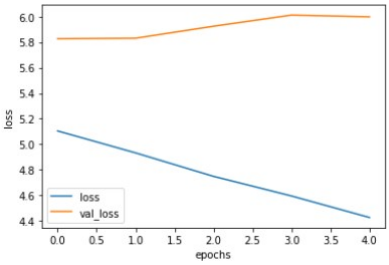
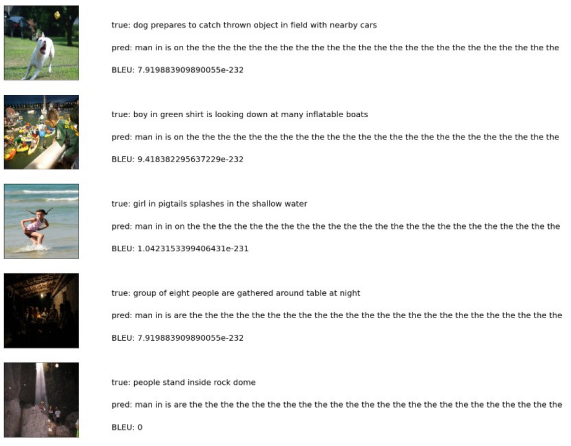


Figure 4 epochs VS val_loss



In the above **Figure 5** shows caption for the images respectively

VI. CONCLUSION

NIC is a comprehensive neural-network system developed to automatically produce coherent english descriptions based on the input images it incorporates a convolutional neural-network to condense images into concise representations followed by a recurrent neural-network responsible for generating corresponding sentences through its training process the model learns to maximize the probability of generating accurate sentences given the input image our experiments conducted on various datasets demonstrate the reliability of NIC both qualitatively wherein the generated sentences display coherence and quantitatively as evaluated by ranking metrics and bleu a standard machine translation metric these findings suggest that as the size of

DataSets	Size of the dataset		
	Training-Data	Validating-Data	Testing-Data
Flickr-8k	6000	1000	1000
Flickr-30k	28000	1000	1000
MSCOCO	82783	40504	40775

image description datasets increases methods like NIC are likely to exhibit performance enhancements moreover exploring the integration of unsupervised data derived from both images and text independently shows potential for further improving image description techniques

VII. REFERENCES.

[1] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences. 232. 01052. 10.1051/matecconf/201823201052.

[2] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.

[3] S. Das, L. Jain and A. Das, "Deep Learning for Military Image Captioning," 2018 21st International Conference on Information Fusion (FUSION), 2018, pp. 2165-2171, doi: 10.23919/ICIF.2018.8455321

[4] GGeetha,T.Kirthigadevi,G GODWIN Ponsam,T.Karthik,M.Safa," Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under licence by IOP Publishing Ltd in Journal of Physics :Conference Series ,Volume 1712, International Conference On Computational Physics in Emerging Technologies(ICCPET) 2020 August 2020,Manglore India in 2015.

[5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014). [6] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recogni-tion and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[6] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.

[7] Chen, Xinlei, and C. Lawrence Zitnick. "Mind's eye: A recurrent visual representation for image caption generation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[8] Feng, Yansong, and Mirella Lapata. "How many words is a picture worth? automatic caption generation for news images." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Com- ´ mon objects in context. arXiv:1405.0312, 2014. 5

[10] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Explain images with multimodal recurrent neural networks. In arXiv:1410.1090, 2014. 2, 6, 7

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In ICLR, 2013. 6, 8

[12] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. C. Berg, K. Yamaguchi, T. L. Berg, K. Stratos, and H. D. III. Midge: Generating image descriptions from computer vision detections. In EACL, 2012. 2

[13] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS, 2011. 2, 5, 6

[14] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: A method for automatic evaluation of machine translation. In ACL, 2002. 4

[15] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139– 147, 2010. 5

[16] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images us-ing 1 million captioned photographs." Advances in neural information processing systems. 2011.