Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

IMAGE CAPTION GENERATOR USING CNN AND LSTM

[1]Mrs. Busani Sravani,[2]S. Sreepragna,[3]R. Madhuri,[4]V. Roja
[1]Assistant Professor,[234]Students
Department Of CSE
Malla Reddy Engineering College for Women

ABSTARCT:

Machine learning is now all the rage in the AI world. We have recently used AI to construct very clever devices with exceptional performance. Deep learning is a subset of machine learning that produces very accurate findings, which in turn indicates very good performance. Apps for picture description make use of deep learning in our study. Providing a description of a picture's content is what image description is all about. Object and action detection in the input picture is the foundation of the notion. When describing images, there are primarily two methods: bottom-up and top-down. Bottom-up methods create captions by combining the information of an input picture. Using different architectures, such as recurrent neural networks, top-down methods provide a semantic representation of an input picture, which is then translated into a caption. One potential advantage of picture description is that it might aid those with visual impairments in comprehending what is shown in online images. What follows is an explanation of the specifics. Looking at the image below, what can you make out?

I. INTRODUCTION

SOCIAL media are platforms for online community-based communication that prioritize user-generated content (UGC), engagement, sharing, and collaboration. Users are able to exchange text, video, and photographs via these mediums. Comment sections and hashtags are common ways for users to amplify the impact of their posts. Additional information on the user's postings and other details may be found in the alternative text (comment, hashtags, etc.). Using data enhanced by social media and other sources, including comments left on YouTube videos, Preece et al. build a Sentinel platform to better comprehend a variety of scenarios. Novel technique for presenting large-scale synthetic social media data is presented by Sagduyu et al. Their technique trains the n-gram model using textual input, namely tweets' hyperlinks and hashtags, to generate themes. Hashtags allow users of various social media platforms to add commentary to the digital information they share, such as Instagram, Facebook, and Twitter. Creators and content providers may apply labeling that makes it

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

simpler for other users to identify their postings by using hashtags, which are often words or nonspaced sentences followed by the sign #. Visual material, such as photos and videos, makes up the bulk of the digital content posted on social networking sites. The difficulty of successfully retrieving photographs from the web and social media is, however, increasing daily. Modern image search engines rely heavily on text descriptions to find relevant results; however, due to shortcomings in these descriptions and an abundance of photos without word annotations, researchers have been devoting more time and energy to developing content-based image retrieval methods. The so-called semantic gap is the meat and potatoes of content-based picture retrieval; although people utilize high-level ideas when searching, content-based retrieval is linked with low-level attributes.

The development of AIA techniques—in which computer systems automatically annotate photographs with information like as captions or keywords—was a response to this issue. The learning-by-example paradigm is perhaps the most used AIA approach. To enable automated annotation of additional (unseen) photos, models are trained on a limited number of training images that have been manually annotated. These models learn the link between image attributes and textual phrases, which are high-level ideas. Good training examples, defined here as accurate and representative image-tagging pairs, are obviously crucial. Instagram in particular is a treasure trove of image-tag combinations found on social media. It is crucial to mine the appropriate ones automatically or semiautomatically so that they may be used as teaching examples.Keep in mind that the majority of the time, the hashtags people use to tag images on social media have nothing to do with the actual content of the images themselves. Instead, they serve multiple purposes, like expressing the user's mood, making the user more discoverable, or starting a new conversation. No more than 25% of Instagram hashtags really represent the image's visual content, according to our earlier study [12]. Also, for the sake of searchability, we've seen a lot of Instagram photos that don't even include a hashtag.

"Stop hashtags" is what we gave them. So, it's necessary to sort the Instagram hashtags according to the image's visual content. We can find the most relevant Instagram hashtags by using a ranking system called Hyperlink-induced topic search (HITS). Jon Kleinberg created the HITS algorithm with the intention of ranking websites. The premise is that a website may serve as a repository for topic-related information and pertinent connections. So, there are two types of websites: those that provide a wealth of information on a certain subject ("authoritative") and those that serve as "hubs" for other

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

related websites. Each website receives a hub and an authoritative value from the HITS algorithm. As mentioned in our earlier work, we have begun to play around with the HITS algorithm in an effort to mine informative Instagram hashtags. In this paper, we expand on that by exploring how the algorithm could be applied in a real-life crowdtagging setting, made possible by the Figure-eight crowdsourcing platform (previously Crowdflower). Moreover, we have raised the maximum amount of annotations to 500 per picture, created bipartite graphs for each image, and computed the annotators' performance on all of those photos. In addition, the suggested strategy is tested against FolkRank to see how well it performs.

II. LITERATURE REVIEW

One approach to automated picture annotation is topic modeling based on Instagram hashtags.

Argyris Argyrou, Stamatios Giannoulakis, and Nicolas Tsapatsoulis are the authors.

The term used to describe the process of automatically tagging digital photographs is "Automatic Image Annotation" (AIA). A large number of current approaches to automated picture annotation use the learning-by-example paradigm. The first and most important stage in these approaches is to construct the training examples, which are picture pairings with associated tags. According to our prior research, Instagram in particular offers a rich supply of hashtags that may be used to build AIA training sets. Unfortunately, our research shows that only 20% of Instagram hashtags really reflect the image's content, therefore a filtering process is necessary to find the right hashtags. In order to forecast the subject of associated Instagram photos, we use topic modeling using Latent Dirichlet Allocation (LDA) in this article. Considering that a subject is made up of related phrases, the suggested approach for identifying the visual theme of an Instagram picture yields a collection of tags that might be useful for training AIA algorithms.

2. Leveraging crowdsourcing to answer multiple-choice questions

Aydin Bahadir Ismail, Yilmaz Yavuz Selim, Li Yaliang, Li Qi, Gao Jing, and Demirbas Murat

We increase the aggregate accuracy of crowdsourced responses to multiple-choice questions by using crowd knowledge and employing lightweight machine learning approaches. We built and released a crowdsourcing system that mimics the "Who wants to be a millionaire?" game show in an effort to

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

improve aggregation algorithms and test them experimentally. After reviewing our data set, which includes over 200,000 responses, we discovered that selecting the most popular answer in the aggregate allows us to properly answer more than 90% of the questions. However, when it comes to the more challenging questions later on in the quiz, this strategy only yields a 60% success rate. We explore new weighted aggregation strategies to aggregate crowdsourced replies in an effort to boost performance on these later, more challenging topics. We demonstrate that we can get an overall average accuracy of 95% and a 15% improvement for the more difficult questions by utilizing weights designed for dependability of participants, which are generated from the confidence of the participants. Our findings provide credence to the idea that crowdsourced question answering systems may be improved with the use of machine learning methods.

3. Crowdsourced classifications of realistic driving scenes: validity and dependability

The need for people to classify massive amounts of recorded visual information is a typical obstacle when processing data from realistic driving scenarios. We tested the hypothesis that crowdsourcing could more accurately classify driving scene elements (such as the presence of other road users, straight road segments, etc.) than a single researcher or small team could using the internet platform CrowdFlower. Two hundred employees from forty-six nations took part in the 1.5-day event. We tested reliability and validity with and without using Gold Test Questions (GTQs), a CrowdFlower technique that allows researchers to insert their own control questions. We discovered that using GTQs greatly improved the external workers' ability to accurately and consistently identify things from driving scenes. With GTQs, the accuracy (i.e., relative to the evaluations of a confederate researcher) on items was 91% at a small size CrowdFlower Job of 48 three-second video segments, compared to 78% without. External workers returned more false positives without GTQs compared to with GTQs, suggesting a difference in bias. We published 12,862 three-second video segments for annotation at a bigger size CrowdFlower Job that only used GTQs. Validating all 1012 classifications at once would be impossible (and counterproductive), so we chose a random selection and found that they all returned 95% accuracy.

4. A review and study of AI to annotate images

The exponential expansion of image data in the last few years has focused a lot of interest on picture annotation. Image annotation's capacity to describe pictures at the semantic level opens up several

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

possibilities in fields related to image analysis and comprehension, including biomedical engineering and urban management. Automatic Image Annotation (AIA) has been proposed since the late 90s as a solution to the problems with manual image annotation. This report presents a comprehensive overview of current AIA methodologies by analyzing 138 publications published during the last 20 years. Here are five ways that AIA approaches are categorized: The first three methods use generative models; the second uses nearest neighbor models; the third uses discriminative models; the fourth uses tag completion models; and the fifth uses deep learning models for picture annotation. Using criteria such as central concept, primary contribution, model framework, computational complexity, computation time, and annotation accuracy, we compare the five AIA techniques. In addition, we provide a synopsis of four standard assessment criteria and five publicly accessible picture datasets that are often used as standards for assessing AIA techniques. Standard evaluation measures and a benchmark dataset are then used to evaluate the performance of a few typical or well-behaved models. We conclude by outlining our thoughts on current and future research trends in AIA and discussing the outstanding concerns and difficulties within the field.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM

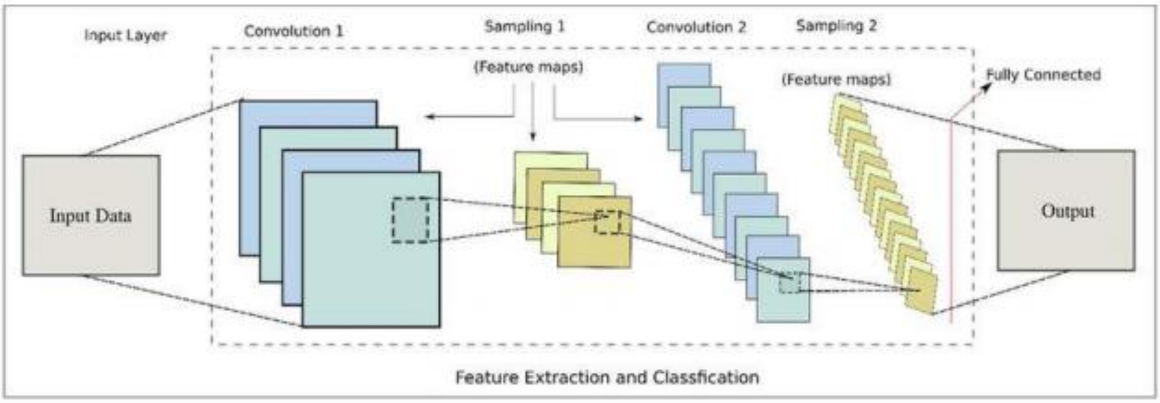
We'll go over the findings from the experiments that used the MSCOCO dataset. A new feature called guiding network has been introduced to the encoder/decoder architecture in their proposed work. The major focus of the guiding network technique is to train a neural network to learn the vector v=g(A), where A is the collection of annotation vectors. It is a significant challenge to provide visual data descriptions in natural language. For a long time, computer vision has been studying it. As a result, sophisticated systems have emerged that integrate visual basic recognizers with formal, structured languages, such as logic systems or And-Or Graphs.The challenge of describing still images using natural language has recently attracted a lot of attention.

PROPOSED SYSTEM

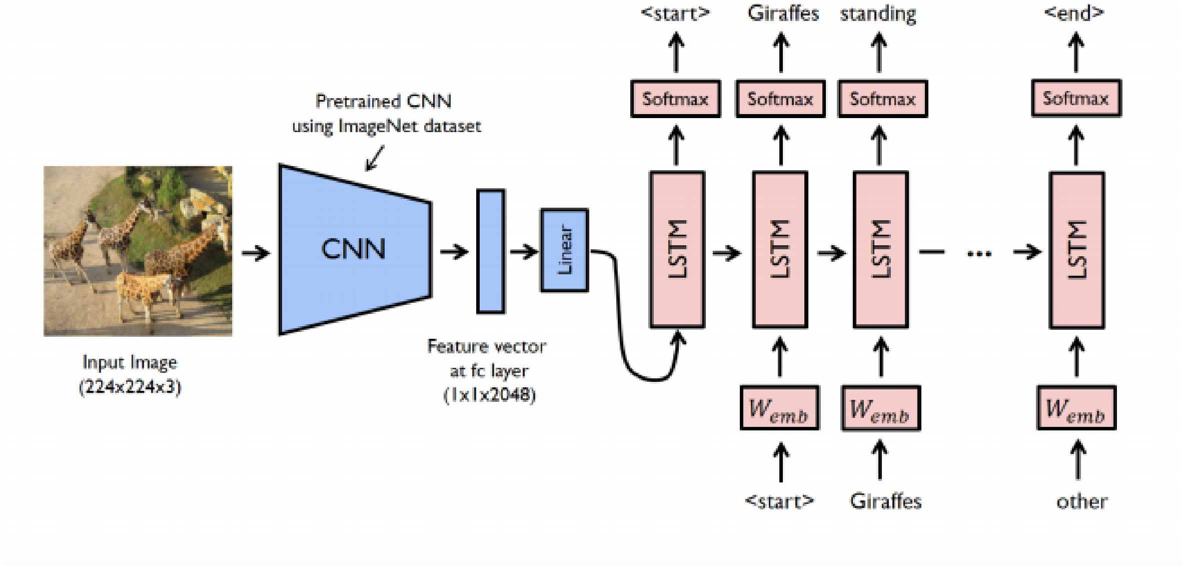
On this page Our objective is to create an image caption generator, and to do this, we employ convolutional neural networks (CNNs) and long short-term memories (LSTMs). Deep learning neural networks may be artificially trained using convolutional neural networks. Classification of images, object identification, picture recognition, and computer vision are some of its many applications.

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

CNN image classifications work by processing incoming images and assigning them to predefined categories, such as "dog," "cat," etc. In order to identify photos, it scans them horizontally and vertically to extract key elements, which it then combines. What exactly is Long Short-Term Memory (LSTM)? The acronym LSTM refers to a specific kind of recurrent neural network (RNN) that excels in solving sequence prediction issues. We may anticipate the next word by looking at the preceding text. By getting above the shortcomings of classical RNN, such as its short-term memory, it has shown its effectiveness. With the use of a forget gate, LSTM can analyze inputs while carrying out important information while discarding irrelevant information. A CNN-RNN model was created by combining these two models.in general The aforementioned top-down picture generating models serve as inspiration for our method. We extract visual picture characteristics using a deep convolutional neural network, and we extract semantic features using the semantic tagging model. Coupling visual characteristics from CNN with semantic features from the tagging model, these features are sent into an LSTM network, which in turn creates captions.



Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800



IV. MODULES

CNN

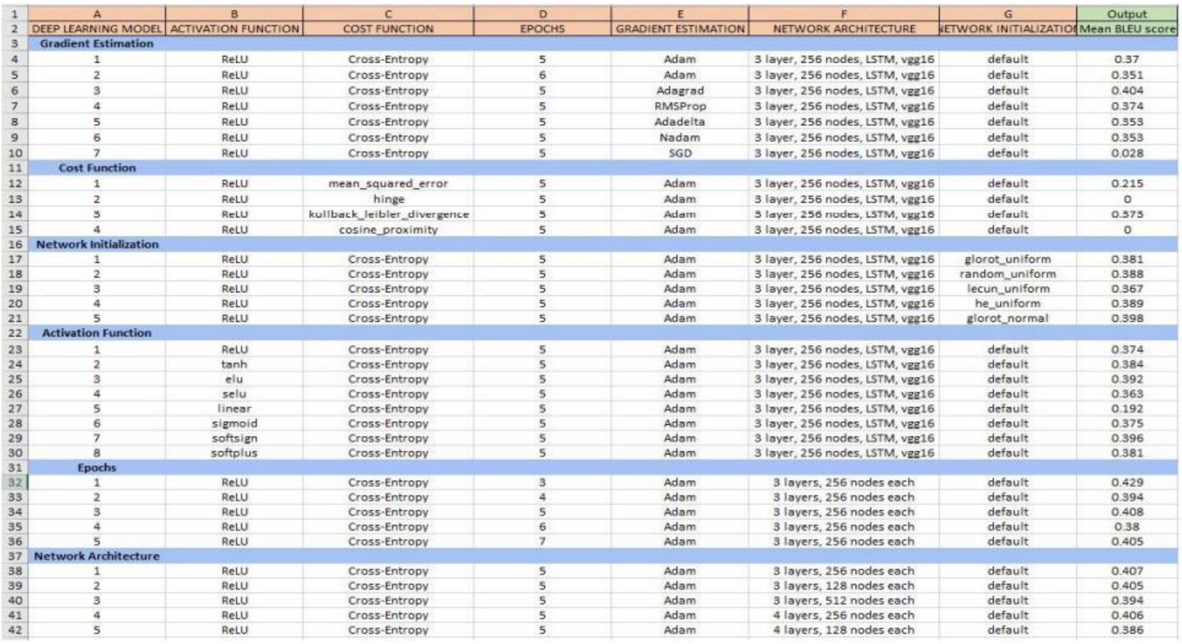
Hidden Markov Model (HMM) Convolutional neural networks are now being used for visual recognition. Multiple convolutional layers make up a CNN. Fully connected layers follow these convolutional layers, similar to a multilayer neural network [14]. The convolutional neural network (CNN) is designed to use the 2D structure of the input picture. To accomplish this, we use various pooling algorithms, adjust the quantity of local connections and linked weights, and the end result is features that are translation invariant. Compared to other networks with the same amount of hidden states, CNN has fewer parameters and is easy to train, which are its main advantages. In this research, we used a Deep Convolutional Neural Network (CNN) for large-scale image recognition called the Visual Group Geometry (VGG) network [15]. You may have it with either 16 or 19 layers. In both the validation and test sets, the classification error values for 16 and 19 layers are quite close to one other, coming in at about 7.4% and 7.3%, respectively. Using this approach, we can define what kinds of images go into making captions.

LSTM

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

Memory for the Short Term (LSTM) A recurrent neural network is used to model the fleeting behavior of a set of entities [17][19]. Due to disappearing and expanding weights or gradients, ordinary RNNs struggle to acquire long-term dynamics [9]. One of the most fundamental components of an LSTM is the memory cell. Over a long period of time, it records the current value. Using gates, one may control how often a cell's state is updated. A variation's representation is the number of links between gates and memory cells. The LSTM block, upon which our model is built, relies on the LSTM that does not have a peephole design. Relationships between gates and long short-term memory cells are as follows:

V. RESULTS



1	A	B	C	D	E	F	G	Output
2	DEEP LEARNING MODEL	ACTIVATION FUNCTION	COST FUNCTION	EPOCHS	GRADIENT ESTIMATION	NETWORK ARCHITECTURE	ETWORK INITIALIZATIO	Mean BLEU score
3	Gradient Estimation							
4	1	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.37
5	2	ReLU	Cross-Entropy	6	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.351
6	3	ReLU	Cross-Entropy	5	Adagrad	3 layer, 256 nodes, LSTM, vgg16	default	0.404
7	4	ReLU	Cross-Entropy	5	RMSProp	3 layer, 256 nodes, LSTM, vgg16	default	0.374
8	5	ReLU	Cross-Entropy	5	Adadelta	3 layer, 256 nodes, LSTM, vgg16	default	0.353
9	6	ReLU	Cross-Entropy	5	Nadam	3 layer, 256 nodes, LSTM, vgg16	default	0.353
10	7	ReLU	Cross-Entropy	5	SGD	3 layer, 256 nodes, LSTM, vgg16	default	0.028
11	Cost Function							
12	1	ReLU	mean_squared_error	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.215
13	2	ReLU	hinge	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0
14	3	ReLU	kullback_leibler_divergence	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.375
15	4	ReLU	cosine_proximity	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0
16	Network Initialization							
17	1	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	glorot_uniform	0.381
18	2	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	random_uniform	0.388
19	3	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	lecun_uniform	0.367
20	4	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	he_uniform	0.389
21	5	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	glorot_normal	0.398
22	Activation Function							
23	1	ReLU	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.374
24	2	tanh	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.384
25	3	elu	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.392
26	4	selu	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.363
27	5	linear	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.192
28	6	sigmoid	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.375
29	7	softsign	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.396
30	8	softplus	Cross-Entropy	5	Adam	3 layer, 256 nodes, LSTM, vgg16	default	0.381
31	Epochs							
32	1	ReLU	Cross-Entropy	3	Adam	3 layers, 256 nodes each	default	0.429
33	2	ReLU	Cross-Entropy	4	Adam	3 layers, 256 nodes each	default	0.394
34	3	ReLU	Cross-Entropy	5	Adam	3 layers, 256 nodes each	default	0.408
35	4	ReLU	Cross-Entropy	6	Adam	3 layers, 256 nodes each	default	0.38
36	5	ReLU	Cross-Entropy	7	Adam	3 layers, 256 nodes each	default	0.405
37	Network Architecture							
38	1	ReLU	Cross-Entropy	5	Adam	3 layers, 256 nodes each	default	0.407
39	2	ReLU	Cross-Entropy	5	Adam	3 layers, 128 nodes each	default	0.405
40	3	ReLU	Cross-Entropy	5	Adam	3 layers, 512 nodes each	default	0.394
41	4	ReLU	Cross-Entropy	5	Adam	4 layers, 256 nodes each	default	0.406
42	5	ReLU	Cross-Entropy	5	Adam	4 layers, 128 nodes each	default	0.386

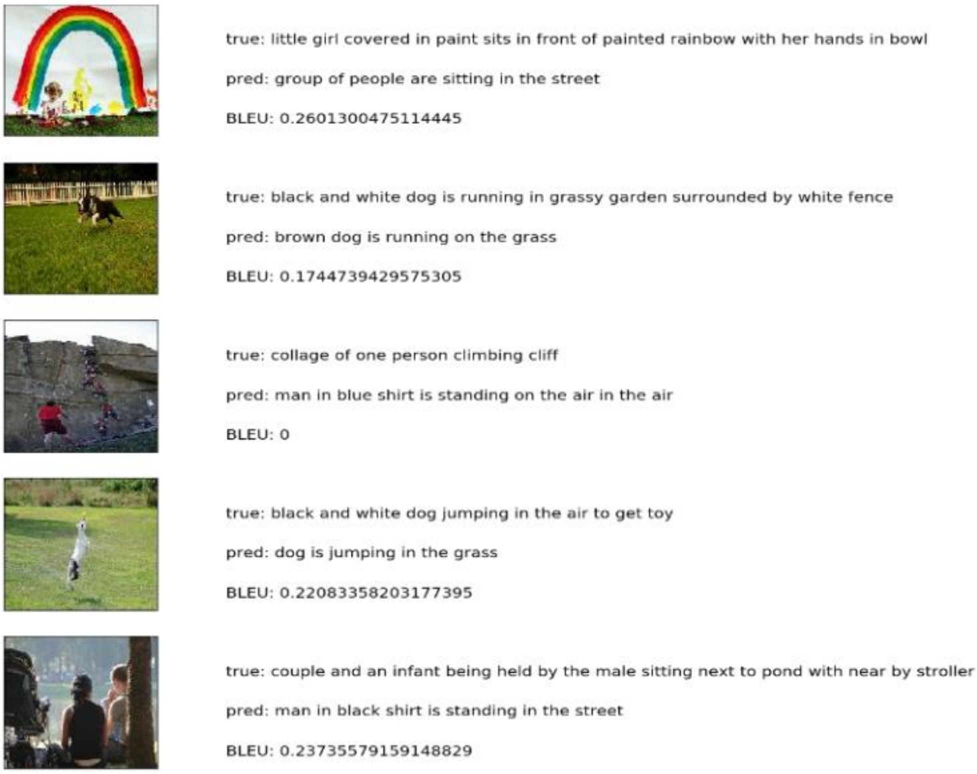
Fig: Model results

Table: BELU Scores

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

No.	Research Works i.e., Models	BELU Scores
1.	LRCN	0.669
2.	NIC	0.277
3.	VSA	0.584
4.	CNN-LSTM	0.681
5.	Our Model	0.398

Some good and bad captions



true: little girl covered in paint sits in front of painted rainbow with her hands in bowl

pred: group of people are sitting in the street

BLEU: 0.2601300475114445

true: black and white dog is running in grassy garden surrounded by white fence

pred: brown dog is running on the grass

BLEU: 0.1744739429575305

true: collage of one person climbing cliff

pred: man in blue shirt is standing on the air in the air

BLEU: 0

true: black and white dog jumping in the air to get toy

pred: dog is jumping in the grass

BLEU: 0.22083358203177395

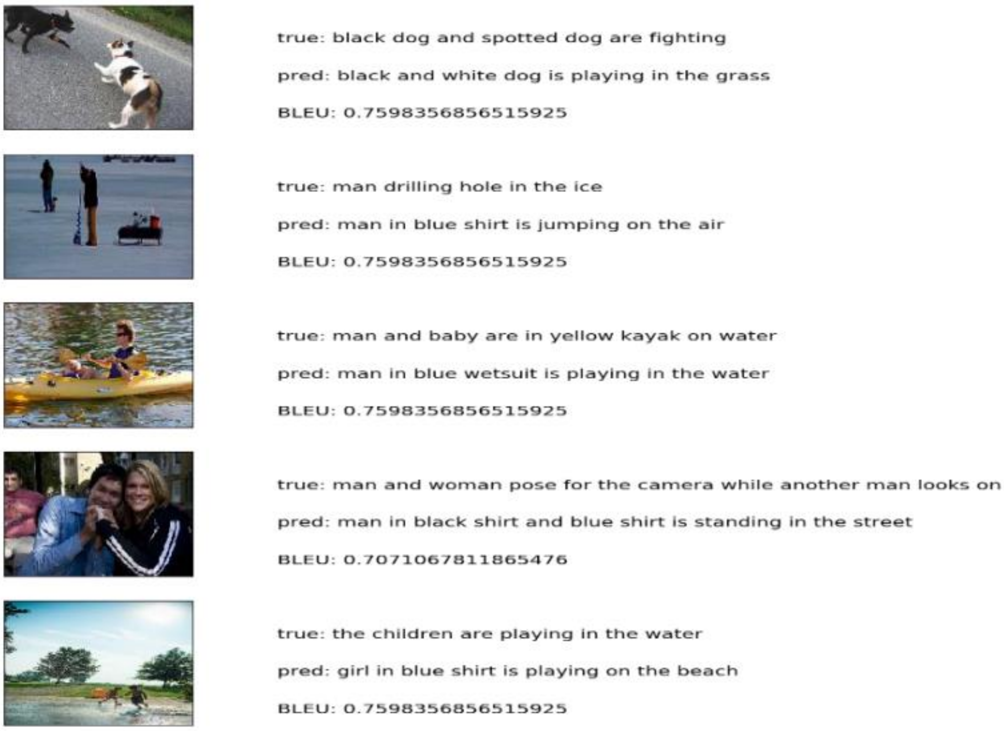
true: couple and an infant being held by the male sitting next to pond with near by stroller

pred: man in black shirt is standing in the street

BLEU: 0.23735579159148829

Fig: Bad Captions

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800



true: black dog and spotted dog are fighting

pred: black and white dog is playing in the grass

BLEU: 0.7598356856515925

true: man drilling hole in the ice

pred: man in blue shirt is jumping on the air

BLEU: 0.7598356856515925

true: man and baby are in yellow kayak on water

pred: man in blue wetsuit is playing in the water

BLEU: 0.7598356856515925

true: man and woman pose for the camera while another man looks on

pred: man in black shirt and blue shirt is standing in the street

BLEU: 0.7071067811865476

true: the children are playing in the water

pred: girl in blue shirt is playing on the beach

BLEU: 0.7598356856515925

Fig: Good Captions

VI. CONCLUSION

Using a dropout keep probability of 75% and a decoder LSTM network with two layers, we ran a thorough hyperparameter search across the CNN-LSTM model architecture and arrived at the best model, which lags behind the state-of-the-art by 3.3 BLEU-4 points and 3.8 CIDEr points. A comprehensive evaluation of the produced metrics, both quantitatively and qualitatively, indicates that the model can appropriately label a broad range of MSCOCO pictures. When people fail to pay close enough attention to details in photos, they often make partial mistakes. For example, a picture of elephants wandering in an enclosure might be mistakenly labeled as "elephants in a field" because the trees in the backdrop led others to believe otherwise. Accordingly, it's possible that this job may benefit from the attention-mechanisms investigated in more recent studies. Our primary original contribution is investigating how out-of-band words influence the LSTM's hitherto unexplored hidden

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

states. We showed that words with comparable semantic closeness (like "plate" and "bowl") when emitted have similar hidden state movements regardless of the prior context, and that differences in hidden state only happen when words with different semantic distances (like "vase" and "food") are emitted.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. [4] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014. [5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014. [6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

[7] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. CoRR, abs/1603.03925, 2016.

[8] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.

Vol. 15 No. 3 (2024):266-277
DOI: https://doi.org/10.61841/turcomat.v15i3.14800

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.

[10] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural Comput., 9(8):1735–1780, November 1997.

[11] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.

[12] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollar, ´ Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. CoRR, abs/1411.4952, 2014.

[13] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. CoRR, abs/1506.07285, 2015.

[14] Alex Graves. Generating sequences with recurrent neural networks. CoRR, abs/1308.0850, 2013.

[15] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. CoRR, abs/1502.04623, 2015.