Journal of Big Data

RESEARCH

Open Access

Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction

Reshmi Sasibhooshan[1*], Suresh Kumaraswamy[2] and Santhoshkumar Sasidharan[3]

*Correspondence:
reshmibhooshan@cet.ac.in

[1] Department of Electronics and Communication Engineering, College of Engineering Trivandrum, Thiruvanathapuram 695016, Kerala, India
[2] Department of Electronics and Communication Engineering, Government Engineering College, Mananthavady, Wayanad 670644, Kerala, India
[3] Department of Electronics and Communication Engineering, Government Engineering College, Painav, Idukki 685603, Kerala, India

Abstract

Automatic caption generation with attention mechanisms aims at generating more descriptive captions containing coarser to finer semantic contents in the image. In this work, we use an encoder-decoder framework employing Wavelet transform based Convolutional Neural Network (WCNN) with two level discrete wavelet decomposition for extracting the visual feature maps highlighting the spatial, spectral and semantic details from the image. The Visual Attention Prediction Network (VAPN) computes both channel and spatial attention for obtaining visually attentive features. In addition to these, local features are also taken into account by considering the contextual spatial relationship between the different objects. The probability of the appropriate word prediction is achieved by combining the aforementioned architecture with Long Short Term Memory (LSTM) decoder network. Experiments are conducted on three benchmark datasets—Flickr8K, Flickr30K and MSCOCO datasets and the evaluation results prove the improved performance of the proposed model with CIDEr score of 124.2.

Keywords: Image captioning, Wavelet transform based Convolutional Neural Network, Visual Attention Prediction, Contextual Spatial Relation Extraction

Introduction

Images are extensively used for conveying enormous amount of information over internet and social media and hence there is an increasing demand for image data analytics for designing efficient information processing systems. This leads to the development of systems with capability to automatically analyze the scenario contained in the image and to express it in meaningful natural language sentences. Image caption generation is an integral part of many useful systems and applications such as visual question answering machines, surveillance video analyzers, video captioning, automatic image retrieval, assistance for visually impaired people, biomedical imaging, robotics and so on. A good captioning system will be capable of highlighting the contextual information in the image similar to human cognitive system. In the recent years, several techniques for automatic caption generation in images have been proposed that can effectively solve many computer vision challenges.

Springer Open

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Basically image captioning is a two step process, which involves a thorough understanding of the visual contents in the image followed by the translation of these information to natural language descriptions. Visual information extraction includes the detection and recognition of objects and also the identification of their relationships. Initially, image captioning is performed using rule based or retrieval based approaches [1, 2]. Later advanced image captioning systems are designed using deep neural architectures that uses a Convolutional Neural Network (CNN) as encoder for visual feature extraction and a Recurrent Neural Network (RNN) as decoder for text generation [3]. Algorithms using Long Short Term Memory (LSTM) [4, 5] and Gated Recurrent Unit (GRU) [6] are introduced to obtain meaningful captions. Inclusion of attention mechanism in the model helps to extract the most relevant objects or regions in the image [7, 8], which can be used for the generation of rich textual descriptions. These networks localize the salient regions in the images and produce improved image captions than previous works. However, they failed to extract accurate contextual information from images, defining the relationship between various objects and between each object and the background, and to project the underlying global and local semantics.

In this work, an image captioning method is proposed that uses discrete wavelet decomposition along with convolutional neural network (WCNN) for extracting the spectral information in addition to the spatial and semantic features of the image. An attempt is made to enhance the visual modelling of the input image by the incorporation of DWT pre-processing stage together with convolutional neural networks that helps to extract some of the distinctive spectral features, which are more predominant in the sub band levels of the image in addition to the spatial, semantic as well as channel details. This helps to include more finer details of each object, for example, the spatial orientation of the objects, colour details etc. In addition to these, it helps to detect the visually salient object/region in the image that draws more attraction similar to human visual system due to its peculiar features with respect to the remaining regions. A Visual Attention Prediction network (VAPN) and Contextual Spatial Relation Extractor (CSE) are employed to extract the fine-grained semantic and contextual spatial relationship existing between the objects, gathered from the feature maps obtained using WCNN. Finally, these details are fed to LSTM network for generating the most relevant captions. The word prediction capability of the LSTM decoder network can be improved through the concatenation of attention based image feature maps and contextual spatial feature maps with the previous hidden state of the language generation model. This enhances the sentence formation greatly, as each word in the generated sentence focuses only on particular spatial locations in the image and its contextual relation with other objects in the image.

The contributions made in this work are:

1 An image caption generation technique incorporating semantic, spatial, spectral and contextual information contained in the image is proposed.
2 A Visual Attention Prediction network is employed to perform atrous convolution on the feature maps generated by WCNN for extracting more semantic contents from the image. Both channel as well as spatial attention are computed from these feature maps.

3 A Contextual Spatial Relation Extractor that utilizes the feature maps generated by WCNN model for predicting region proposals to detect spatial relationship between different objects in the image.

The performance of the method is evaluated using three benchmark datasets—Flickr8K, Flickr30K and Microsoft COCO datasets and comparison is done with the existing state-of-the-art methods using the evaluation metrics—BLEU@N, METEOR, ROUGE-L and CIDEr.

The organization of the paper is as follows: First a brief descriptions about the previous works in image captioning, which is followed by the proposed model architecture and detailed experiments and results. Finally the conclusion of the work is also provided.

Related works

In order to generate an appropriate textual description, it is necessary to have a better understanding about the spatial and semantic contents of the image. As mentioned above, the initial attempts of image caption generation are carried out by extracting the visual features of the image using conditional random fields (CRFs) [9, 10] and translating these features to text using sentence template or combinatorial optimization algorithms [1, 11, 12]. Later retrieval based approaches are used to produce captions that involves the process of retrieving one or a set of sentences from a pre-specified sentence pool based on visual similarities [2, 13]. The evolution of deep neural network architectures helps to have visual and natural language modelling in a superior manner by generating more meaningful descriptions of the image.

The inclusion of additional attention mechanisms in the encoder—decoder framework helps to extract more appropriate semantic information from the image and thereby creating captions that look similar to human generated captions [14–16]. An encoder–decoder model capable of detecting dynamically attentive salient regions in the image during the creation of textual descriptions is proposed by Xu et al. [17]. Yang et al. [18] proposed a review network with enhanced global modelling capabilities for image caption generation. A Fusion-based Recurrent Multi-Modal (FRMM) model consisting of CNN-LSTM framework together with a fusion stage generates captions by combining visual and textual modality outputs [19]. A better comprehensive description can be generated using a Recurrent Fusion Network (RFNet) [20] or by modifying the CNN-LSTM architecture by incorporating semantic attribute based features [21] and context-word features [22]. Attempts are also made by simultaneously introducing a dual attention strategy in both visual as well as textual information to improve the image understanding [23, 24]. Better descriptions can be produced using multimodal RNN framework that makes use of inferred alignments between segments of the image and sentences describing them [25]. To make mandatory correspondence between descriptive text words and image regions effective, Deng et al. proposed a Dense network and adaptive attention technique [26]. A multitask learning method through a dual learning mechanism for cross-domain image captioning is proposed in [27]. It uses reinforced learning algorithm to acquire highly rewarded captions. Attempts for better caption generation has also been done with the development of algorithms with multi-gate attention networks [28], CaptionNet model with reduced dependency on previously predicted words

[29], context aware visual policy networks [30] and adaptive attention mechanisms [31–33]. A new positional encoding scheme is proposed for enhancing the object features in the image, which greatly improves the attention information for generating more enriched captions is presented in [34]. A novel technique to understand the image and language modality configuration for achieving better semantic extraction from images using anchor points are presented in [35]. Visual language modelling can also be accomplished using a Task-Agnostic and Modality Agnostic sequence-to-sequence learning framework [36]. Another image captioning technique employs a configuration that utilizes a set of memory vectors together with meshed connectivity between encoding and decoding sections of the transformer model [37]. A scaling rule for image captioning is proposed in [38] by introducing a LargE-scale iMage captiONer (LEMON) dataset, which is capable of identifying long-tail visual concepts even in a zero shot mechanism. [39] presents a novel methodology for vision language task that is pre-trained on much larger text-image corpora and is able to collect much better visual features or concepts.

Even though these approaches perform well in image caption generation, they still lack the inclusion of fine grained semantic details as well as contextual spatial relationship between different objects in the image, which need to be improved further with enhanced network architectures. Also, due to the presence of multiple objects in the image, it is very essential that the visual contextual relationship between objects must be in correct sequential order to generate a caption with which the contents in the image is better represented. This can be solved by introducing attention mechanisms and considering the spatial relationship between objects in the model.

Model architecture

The model consists of an encoder-decoder framework with VAPN, that converts an input image I to a sequence of encoded words, $W = [w_1, w_2,...w_L]$, with $w_i \in \mathbb{R}^N$, describing the image, where L is the length of the generated caption and N is the vocabulary size. The detailed architecture is presented in Fig. 1. The encoder consists of the WCNN model that incorporates two levels of discrete wavelet decomposition combined with CNN layers to obtain the visual features of the image. The features maps F_1 and F_2 obtained from the CNN layers are bilinearly downsampled and are concatenated together with F_3 and F_4 to produce a combined feature map, F_{in} of size, 32×32×960. This is then given to the VAPN for obtaining attention based feature maps that highlights the semantic details in I by exploiting channel as well spatial attention. In order to extract the contextual spatial relationship between the objects in I, the feature map of level L_4 of WCNN, F_4, is given to the CSE network as shown in Fig. 1. The contextual spatial feature map, F_{cse}, generated by the CSE network is concatenated with the attention based feature map, F_{Att}, to produce F_o and is provided to the language generation stage consisting of LSTM decoder network [40]. The detailed description of the technique is presented below.

Wavelet transform based CNN model

The image is resized to 256×256 and is split up into R, G and B components. Then each of these planes is decomposed into details and approximations using low pass and high pass discrete wavelet filters. In this work, two levels of discrete wavelet decomposition are used that results in the formation of *LL*, *LH*, *HL* and *HH* sub-bands, where *L* and *H*

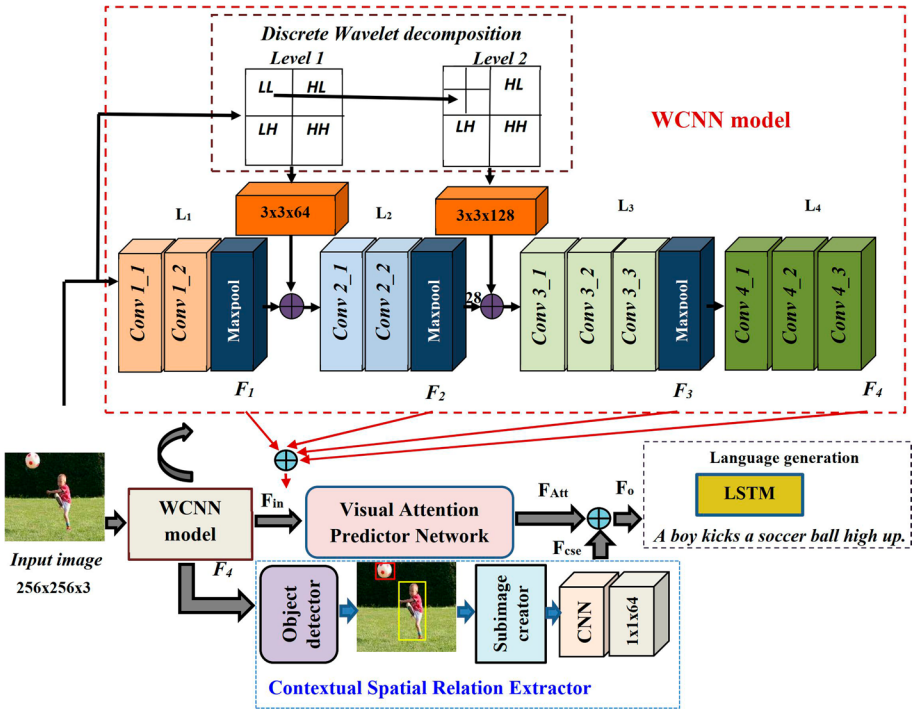


Fig. 1 Proposed model architecture

Table 1 Details of various convolutional layers in WCNN model

Sl. No.	Level	Name of convolutional layers	Kernel size/No. of filters	Output size
1	L1	*Conv 1_1*	3x3/64	256x256x64
2		*Conv 1_2*	3x3/64	256x256x64
3		**Maxpool1**	2x2/64/*stride 2*	128x128x64
4	L2	*Conv 2_1*	3x3/128	128x128x128
5		*Conv 2_2*	3x3/128	128x128x128
6		**Maxpool2**	2x2/128/*stride 2*	64x64x128
7	L3	*Conv 3_1*	5x5/256	64x64x256
8		*Conv 3_2*	5x5/256	64x64x256
9		*Conv 3_3*	5x5/256	64x64x256
10		**Maxpool3**	2x2/256/*stride 2*	32x32x256
11	L4	*Conv 4_1*	7x7/512	32x32x512
12		*Conv 4_2*	7x7/512	32x32x512
13		*Conv 4_3*	7x7/512	32x32x512

represents the low frequency and high frequency components of the input. In the second level decomposition, only the approximations (*LL*) of R, G and B will be further decomposed for each of the three components. These components are then stacked together at each level and are concatenated with the outputs of the first two levels of CNN having four levels (L_1 to L_4) consisting of multiple convolutional layers and pooling layers of kernel size 2×2 with stride 2 as shown in Fig. 1. The details of various convolution layers are given in Tables 1 and 2.

The image feature maps F_1 through F_4 so obtained are then given to the VAPN network for extracting the inter-spatial as well as inter-channel relationship between them.

Visual attention predictor network

To have good insight of the visual features of the image for acquiring the most relevant caption, it is necessary to extract out the semantic features from the input image feature maps, F_1 through F_4. The size, shape, texture features etc. of the objects vary in an image, which causes difficulties in its identification or recognition. To tackle these situations, atrous convolutional network with multi-receptive field filters are employed, which extracts out more semantic details from the image as these filters are capable of delivering a wider field of view at the same computational cost. In this work, four multi-receptive filters—one having 64 filters of kernel size 1×1, and the rest three of kernel size 3×3 each with 64 filters of dilation rates 3, 5 and 7, respectively, are used. The combined feature map, F_{in}, is subjected to atrous convolution and the resultant enhanced feature map of size 32×32×256, is given to a sequential combination of channel attention (CA) network and spatial attention (SA) network as illustrated in Fig. 2. The CA network examines the relationship between channels and more weight will be assigned to those channel having attentive regions to generate the refined channel attention map, whereas the SA network generates a spatial attention map by considering the spatial details between various feature maps.

For more accurate descriptions, the weights of channel attention and spatial attention, W_C and W_S are computed by considering $h_{t-1} \in \mathbb{R}^D$, the hidden state of LSTM memory at $(t-1)$ time step [14]. Here D represents the hidden state dimension. This mechanism helps to include more contextual information in the image during caption generation.

In the channel-wise attention network, the feature map of $F_{in} \in \mathbb{R}^{H \times W \times C}$ is first average pooled channel-wise to obtain a channel feature vector, $V_C \in \mathbb{R}^C$, where H, W and C represent the height, width and total number of channels of the feature map, respectively.

$$V_C = [v_1, v_2 ... v_c] \tag{1}$$

Then channel attention weights, W_C can be computed as,

$$X_C = tanh((W_C^{'} \odot V_C + b_C) \oplus W_{hc} h_{t-1}) \tag{2}$$

$$W_C = softmax((W_o X_C + b_o) \tag{3}$$

where $W_C^{'} \in \mathbb{R}^K$, $W_{hc} \in \mathbb{R}^{K \times D}$ and $W_o \in \mathbb{R}^K$ are the transformation matrices with K denoting the common mapping space dimension of image feature map and hidden state of the LSTM network, $\odot$ and $\oplus$ represent the element-wise multiplication and addition of

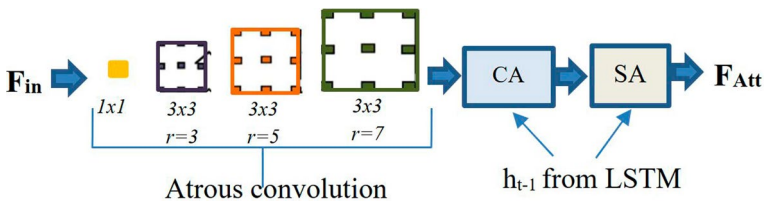


Fig. 2 Structure of VAPN

vectors, $b_C \in \mathbb{R}^K$ and $b_o \in \mathbb{R}^1$ denote the bias terms. Then $\mathbf{F}_{in}$ is element-wise multiplied with W_C and the resultant channel refined feature map, $\mathbf{F}_{ch}$, is given to the spatial attention network.

In the spatial attention network, the weights, W_S, can be calculated with flattened $\mathbf{F}_{ch}$, as follows,

$$X_S = tanh((W_S^{'} \odot \mathbf{F}_{ch} + b_S) \oplus W_{hs} h_{t-1}) \tag{4}$$

$$W_S = softmax(W_o^{'} X_C + b_o^{'}) \tag{5}$$

where $W_S^{'} \in \mathbb{R}^{K \times C}$, $W_{hs} \in \mathbb{R}^{K \times D}$ and $W_o^{'} \in \mathbb{R}^K$ are the transformation matrices, $b_S \in \mathbb{R}^K$ and $b_o^{'} \in \mathbb{R}^1$ denote the bias terms. After this, $\mathbf{F}_{in}$ is element-wise multiplied with W_S to obtain the attention feature map, $\mathbf{F}_{Att}$, highlighting the semantic details in the image.

Contextual spatial relation extractor

Rich image captioning can be achieved only through the exploitation of contextual spatial relation between the different objects in the image in addition to the semantic details. For this, first the regions occupied by the various objects are identified using a network configuration similar to Faster RCNN [41] incorporating WCNN model and RPN along with classifier and regression layers for creating bounding boxes. The feature map F_4 from level L_4 of WCNN model is given as input to the Region Proposal Network (RPN) for finding out the object regions in the image. Then these objects are paired and different sub images containing each of the identified object pairs are created and resized to 32×32 for uniformity. Each of these sub images are given to the CNN layers with three sets of 64 filters, each with receptive field of 3×3, to generate the features describing spatial relations between the object pairs. The spatial relation feature maps of each object pair is then stacked together and is given to 1×1×64 convolution layer to form the contextual spatial relation feature map, F_{cse}, of size 32×32×64, which is further concatenated with F_{Att} to obtain F_o and is given to the LSTM for the generation of the next caption word.

$$h_t = LSTM(h_{t-1}, F_o, y_{t-1}) \tag{6}$$

The prediction of the output word at time step t, is nothing but the probability of selecting suitable word from the pre-defined dictionary containing all the caption words.

$$y_t = softmax(h_t, y_{t-1}) \tag{7}$$

The model architecture is trained to optimize the categorical cross entropy loss, L_{CE}, given by

$$L_{CE}(\theta) = -\sum_{t=1}^{T} log(p_\theta(y_t \mid y_{1:t-1})) \tag{8}$$

where $y_{1:t-1}$ represents the ground truth sequence and θ represents the parameters. The self-critical training strategy [42] is used in this method to solve the exposure bias problem during the optimization with cross-entropy loss alone. Initially, the model is trained

using cross-entropy loss and is further optimized in a reinforced learning technique with self-critical loss for achieving the best CIDEr score as reward on validation set.

Experiments and results

The detailed description of the datasets used and the performance evaluation of the proposed method are presented in this section. Both quantitative as well as qualitative analysis of the method is carried out and is compared with the existing state-of-the-art methods.

Datasets and performance evaluation metrics used

The experiments are conducted on three benchmark datasets: (1) Flickr8K (8,000 images) [43], (2) Flickr30K (31,000 images) [44] and (3) Microsoft COCO dataset (82,783 images in training set, 40,504 images in validation set and 40,775 images in test set) [45]. All the images are associated with five sentences. Among these for the Flickr8K dataset, we have selected 6,000 images for training, 1,000 images for validation and 1,000 images for testing as per the official split of the dataset. For fair comparison with previous works, the publicly available Karpathy split for Flick30K and Microsoft COCO datasets have been adopted [46]. As per this, the Flickr30K dataset split is set to 29,000 images for training, 1,000 images for validation and 1,000 images for testing. In the case of MSCOCO dataset, the split is set as 5,000 images for validation, 5,000 images for testing and all others for training. The textual descriptions are pre-processed with the publicly available code in *https://github.com/karpathy/neuraltalk*, so that all the captions are converted to lowercase and the non-alphanumeric characters are discarded. Each caption length is limited to 40 and those words whose occurrence is less than five times in the ground truth sentences are removed from the vocabulary.

The performance evaluation of the proposed captioning model is done using the evaluation metrics—Bilingual Evaluation Understudy (BLEU@N (for N=1,2,3,4)) [47], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [48], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [49] and Consensus-based Image Description Evaluation (CIDEr) [50] denoted as B@N, MT, R and CD, respectively. These metrics measures the consistency between n-gram occurrences in the caption generated and the ground-truth descriptions. It should be noted that the fair comparison of the results are reported by optimizing the methods with cross entropy loss.

Implementation details

To encode the input image to visual features, the Wavelet transform based CNN model with Birothogonal 1.5 wavelet (bior 1.5) [51] pretrained using ImageNet dataset [52] is used. For the modified Faster R-CNN configuration used in the Contextual spatial relation extractor, we have used an IoU threshold of 0.8 for region proposal suppression, and 0.2 for object class suppression, respectively. The LSTM employed in the decoding section is having a hidden state dimension set as 512, respectively. The caption generation process terminates with the occurrence of a special END token or if predefined max sentence length is reached. Batch normalization is applied to all convolutional layers with decay set as 0.9 and epsilon as 1e-5. Optimization of the model is done using ADAM optimizer [53] with an initial learning rate of 4e-3. The exponential decay rates

for the first and second moment estimates are chosen as (0.8, 0.999). The mini batch size for the Flickr8K, Flickr30K and MSCOCO dataset are taken as 16, 32 and 64. The self critical training strategy is employed in the implementation, where the model is trained initially for 50 epochs with the cross-entropy loss and it is further fine tuned with 15 epochs using the self-critical loss for achieving the best CD score on validation set. To avoid overfitting, dropout rate is set as 0.2 and L2 regularization with weight decay value set as 0.001. For the word vector representation of each word, 300 dimensional GloVe word embeddings [54] pre-trained on a large-scale corpus is employed in this work. For the selection of best caption, BeamSearch strategy [3] is adopted with a beam size of 5, which selects best caption from few selected candidates. The proposed image captioning framework is implemented using TensorFlow 2.3. To train the proposed method, we used Nvidia Tesla V100 with 16GB with 5120 CUDA cores.

Analysis for the selection of appropriate mother wavelet

The choice of the mother wavelet has been done by analyzing the performance of the proposed method on four different wavelet families—Daubechies wavelets (dbN), biorthogonal Wavelets (biorNr.Nd), Coiflets (CoifN) and Symlets (SymN), where N represents number of vanishing moments, Nr and Nd denotes the number of vanishing moments in the reconstruction and decomposition filters, respectively. Table 2 gives the detailed performance results of the proposed image captioning method for Flickr8K, Flickr30K and MSCOCO datasets. Here a model consisting of WCNN with two-level DWT decomposition, VAPN and LSTM alone is considered. The contextual spatial relation extractor is not taken into account for this experimentation. In this, the baseline method (BM) is a model similar to the one described above without DWT decomposition. From Table 2, it is evident that bior1.5 is giving good results compared to other wavelets used in the experimentation. This is because it employs separate scaling and wavelet functions for decomposition and reconstruction purposes. Biorthogonal wavelets exhibit the property of linear phase and allow additional degrees of freedom when compared to orthogonal wavelets.

Table 2 Performance results of the proposed image captioning method for different mother wavelets. Here BM denotes the baseline method

Mother wavelet	Flickr8K		Flickr30K		MSCOCO	
	B@4	CD	B@4	CD	B@4	CD
BM	24.43	58.31	23.68	57.89	35.78	118.02
db1	25.77	59.37	24.87	58.91	36.57	119.84
db4	25.86	59.56	25.01	59.14	36.82	119.95
bior1.5	**26.34**	**60.58**	25.30	**60.13**	**37.14**	**120.41**
bior2.4	26.18	60.52	**25.32**	60.02	37.01	120.16
bior3.5	26.04	60.19	25.03	59.84	36.89	120.03
bior5.5	25.85	59.92	24.84	59.77	36.77	119.98
Coif2	25.96	59.77	24.97	59.52	36.79	119.64
Coif5	26.08	59.63	24.82	59.03	36.62	119.58
Sym2	25.81	59.68	24.73	58.78	36.81	119.80
Sym4	24.97	59.72	24.61	58.65	36.73	119.63

The highest values for each of the metrics are given in bold

Here for the Flickr30K dataset bior2.4 secures highest B@4 score of about 25.32 but bior1.5 scores better value for CD compared to the other wavelets. For Flickr30K and MSCOCO datasets, bior1.5 achieves better B@4 and CD scores of about 26.34 and 60.58 and about 37.14 and 120.41 for the MSCOCO dataset. Hence bior1.5 wavelet is chosen as the mother wavelet for experimentation using these datasets.

Analysis for the choice of DWT decomposition levels

For analysing the performance of the model with different DWT decomposition levels, experiments are carried out with the same baseline model as used for finding out the best mother wavelet. Here bior1.5 wavelet are employed. The detailed ablation study regarding the results obtained for various DWT decomposition levels for the MSCOCO dataset are presented in Table 3.

The model with two level DWT decomposition exhibits better performance compared to 1-level DWT decomposition scoring an improvement in the B@4, MT and CD values of about 1.26%, 1.79% and 1.18%, respectively, for the MSCOCO dataset. The method also shows little performance improvements with the use of three level decomposition stages. Hence considering the computational complexity, method with two level decomposition is preferred in the proposed work.

Quantitative analysis

The quantitative evaluation of the proposed method for the MSCOCO dataset is carried out using the above mentioned metrics and the performance comparison is done with fifteen state-of-the-art methods in image captioning. These models include Deep VS [25], emb-gLSTM [5], Soft and Hard attention [17], ATT [55], SCA-CNN [14], LSTM-A [56], Up-down [7], SCST [42], RFNet [20], GCN-LSTM [57], avtmNet [58], ERNN [59], Tri-LSTM [60] and TDA+GLD [61].

Table 4 shows the evaluation results for the comparative study on MSCOCO 'Karpathy' test split. From Table 4, it is evident that the proposed method outperforms the state-of-the-art methods with good CD score of 124.2. It acquires a relative improvement of about 0.9 %, 0.5 %, 0.2 % and 0.7 % in B@4, MT, R and CD score, respectively, compared to Tri-LSTM method [60]. This is because the image feature maps of the proposed model includes spectral information also along with spatial as well as semantic details compared to the other methods by the inclusion of discrete wavelet decomposition in the CNN model, which helps to extract fine grained information during object detection. Also it takes into account of contextual spatial relationship between objects in the image and exploits both spatial and channel-wise attention of the enhanced feature

Table 3 Performance results of the proposed method for different number of decomposition levels on MSCOCO dataset

Decomposition levels	MSCOCO		
	B@4	MT	CD
1-level	52.87	35.14	90.39
2-level	53.64	36.53	91.71
3-level	53.69	36.90	91.89

Table 4 Performance comparison of the proposed method on MSCOCO dataset. (-) indicates 'metric is not reported'. The best two models having larger values of the metrics are shown in red and green

Method	B@1	B@2	B@3	B@4	MT	R	CD
Deep VS [25]	62.5	45.0	32.1	23.0	19.5	-	66.0
emb-gLSTM [5]	67.0	49.1	35.8	26.4	22.74	-	81.25
Soft attn [17]	70.7	49.2	34.4	24.3	23.9	-	-
Hard attn [17]	71.8	50.4	35.7	25.0	23.04	-	-
ATT [55]	70.9	53.7	40.2	30.4	24.3	-	-
SCA-CNN [14]	71.9	54.8	41.1	31.1	25.0	-	-
LSTM-A [56]	75.4	–	–	35.2	26.9	55.8	108.8
Up-down [7]	77.2	–	–	36.2	27.0	56.4	113.5
SCST [42]	–	–	–	34.2	26.7	55.7	114.0
RFNet [20]	76.4	60.4	46.6	35.8	27.4	56.5	112.5
GCN-LSTM [57]	77.4	–	–	37.1	28.1	57.2	117.1
avtmNet [58]	–	–	–	33.2	27.3	56.7	112.6
ERNN [59]	73.2	56.9	42.9	32.2	25.2	-	101.4
Tri-LSTM [62]	–	–	–	**37.3**	**28.4**	**58.1**	**123.5**
TDA+GLD [61]	*78.8*	*62.6*	**48.0**	36.1	27.8	57.1	121.1
Ours	**78.5**	**62.0**	*49.1*	*38.2*	*28.9*	*58.3*	*124.2*

Table 5 Performance comparison of the proposed method on Flickr8K dataset. The best two models having larger values of the metrics are shown in red and green

Method	B@1	B@2	B@3	B@4	MT
Deep VS [25]	57.9	38.3	24.5	16.0	-
emb-gLSTM [5]	64.7	45.9	31.8	21.2	20.6
Soft attn [17]	67.0	44.8	29.9	19.5	18.9
Hard attn [17]	67.0	45.7	31.4	21.3	20.3
SCA-CNN [14]	**68.2**	**49.6**	**35.9**	**25.8**	**22.4**
Ours	*70.5*	*50.2*	*37.3*	*28.6*	*24.5*

maps resulted from atrous convolution employing multi-receptive field filters that are capable of locating visually salient objects of varying shapes, scales and sizes.

The comparison results of the proposed method for the Flickr8K and Flickr30K datasets are tabulated in Tables 5, 6. From Table 5, it is evident that the method outperforms the state-of-the-art methods with an improvement of 2.3%, 2.8% and 2.1% on B@1, B@4 and MT compared to SCA-CNN [14] for Flickr8K dataset. Also the proposed method scores an improvement of 2.4 % and 0.9 % on B@4 and MT in comparison with the avtmNet [58] method for Flickr30K dataset. It also secures a good CD value of about 67.3.

Qualitative results

Figure 3 shows the captions generated by the proposed image captioning algorithm and the baseline approach for few sample images. Here the baseline method utilizes a VGG-16 network instead of WCNN network, together with the attention mechanism and LSTM configuration. The CSE network is not incorporated in the baseline method. For the qualitative study, images of simple scenes, complex scenes with multiple objects

Table 6 Performance comparison of the proposed method on Flickr30K dataset. The best two models having larger values of the metrics are shown in red and green

Method	B@1	B@2	B@3	B@4	MT	CD
Deep VS [25]	57.3	36.9	24.0	15.7	15.3	
emb-gLSTM [5]	64.6	44.6	30.5	20.6	17.9	-
Soft attn [17]	66.7	43.4	28.8	19.1	18.5	-
Hard attn [17]	66.9	43.9	29.6	19.9	18.5	-
ATT [55]	64.7	46.0	32.4	23.0	18.9	-
SCA-CNN [14]	**66.2**	**46.8**	**32.5**	22.3	19.5	-
avtmNet [58]	–	–	–	**24.8**	**20.8**	**59.8**
Ours	*70.1*	*49.4*	*35.8*	*27.2*	*21.7*	*67.3*

Fig. 3 Comparison of the captions generated by the baseline approach and the proposed method for few samples images. Here **GT** represents the ground truth sentence

and also those having objects that is to be specified clearly with colour as well as their count are chosen. For the images 1 and 4, our method successfully identifies the number of persons in the image as *'three'*. The proposed method generates more enriched captions for the images 1, 2, 5 and 6, by detecting *'wooden benches'*, *'water fountains'*, *'red ropes'* and *'yellow guitar'*. Image 3 consists of more complex scene than others with more number of objects, for which the method is able to distinguish *'five men'* and *'bicycles'* clearly along with the extract of relationship between them as *'standing near'* in a better manner. Thus from Fig. 3, it is evident that the use of attention mechnaism in WCNN

structure and CSE network helps to include more finer details in the image, highlighting both spatial, semantic as well as contextual relationship between the various objects thereby generating better textual descriptions of the image.

Figure 4 shows a few failure captions generated by the proposed captioning method. The method is unable to extract the semantics from certain images due to incorrect detection and recognition of objects and also the contextual relation between them. For images 1 and 3, the objects are not distinguished from the background and in image 6, the activity is wrongly interpreted as '*walking*'. In image 2, the object '*remote*' is identified as '*phone*' and hence the method fails to recognize the activity of that complex scene. Similar error happened in the case of image 5, with the detection of '*rope*' as some fishing tool. Image 4 is formed from a group of photos combined together highlighting the various moves and techniques of tennis players. In this situations, the method fails to extract the semantics and contextual information of the image generating a caption as '*Men **dancing** with tennis racquet*'.

Ablation study

To evaluate the performance of the combination of various stages in the model, an ablation study is conducted. The evaluation results of various configurations of the proposed model with MSCOCO dataset for cross entropy loss and with self critical loss with CD score optimization are given in Table 7. Experimental studies are conducted initially by using WCNN model as the encoder network together with atrous convolution to extract

Fig. 4 Few samples of failure captions generated by the proposed method. The descriptions given in blue and red represents the ground truth and generated captions, respectively

the enhanced feature maps, which are then directly fed to the LSTM network for translation. Then attention networks are included in different combinations to the enhanced feature maps, which generates enriched textual descriptions. At first, SA network alone was incorporated by considering the h_{t-1} hidden state of the LSTM that highlights the spatial locations according to the generation of words to form the descriptions. This improved the B@4 and CD score by about 0.8% and 1.6% for cross entropy loss and about 1.3% and 1.4% for self-critical loss, respectively. Further CA network is introduced and we tried both the sequential combinations with CA and SA networks: (SA+CA) and (CA+SA). The (CA+SA) combinational network proved to be better than the other by providing B@4 and CD score of about 37.1 and 120.4 for self critical loss as shown in Table 7. Finally, the CSE network is also employed in the network, which enhances the B@4 and CD score by 38.2 and 124.2.

Figure 5 illustrates the captions generated by the proposed method for three sample images with their annotated ground-truth sentences. From the generated captions, it is evident that the proposed method is able to understand the visual concepts in the image in a superior manner and is successful in generating more suitable captions, reflecting the underlying semantics.

Conclusion

In this work, a deep neural architecture for image caption generation using encoder-decoder framework along with VAPN and CSE network is proposed. The inclusion of wavelet decomposition in the convolutional neural network extracts spatial, semantic and spectral information from the input image, which along with atrous convolution predicts the most salient regions in it. Rich image captions are obtained by employing spatial as well as channel-wise attention in the feature maps provided by the VAPN and also by considering the contextual spatial relationship between the objects in the image using CSE network. The experiment is conducted with three benchmark datasets, Flickr8K, Flickr16K and MSCOCO using BLEU, METEOR and CIDEr performance metrics. It achieved a good B@4, MT and CD score of 38.2, 28.9 and 124.2, respectively, for MSCOCO dataset. This proves its effectiveness. Since contextual information are utilised, finer semantics can be included in the generated captions. In future works, instead of LSTM, more advanced transformer network can be considered for achieving better CD score and finer captions. By the incorporation of temporal attention, the work can be extended for caption generation in videos.

Table 7 Results of ablation study conducted on MSCOCO dataset

Configuration	Cross-Entropy loss		Self-Critical loss	
	B@4	CD	B@4	CD
WCNN+atr+LSTM	33.1	109.2	34.4	116.5
WCNN+atr+SA+LSTM	33.9	110.8	35.7	117.9
WCNN+atr+SA+CA+LSTM	35.2	112.7	36.3	119.0
WCNN+atr+CA+SA+LSTM	35.9	113.4	37.1	120.4
WCNN+atr+CA+SA+CSE+LSTM	37.5	116.9	38.2	124.2

Here *atr* atrous convolution, *SA* spatial attention, *CA* channel attention

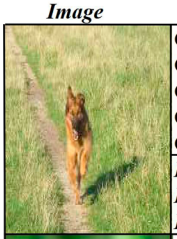
Image	*Captions*
	**GT1:** A brown dog runs along a path in the grass **GT2:** Dog running in field **GT3:** Dog running on narrow dirt path **GT4:** A brown dog is running in a glassy plain **GT5:** The dog is running along a path that has been made through the uncut grass ***D1 : A dog is running*** ***D2 : A brown dog runs along a path*** ***D3 : A brown dog runs along a grassy path***
	**GT1:** The bird leans over to a small piece of food **GT2:** A small pale bird bends down to examine a crumb **GT3:** A small white bird looks at a small object **GT4:** A bird eating **GT5:** A tan bird stands on a ledge about to eat something ***D1 : A bird looking down*** ***D2 : A small bird looking down*** ***D3 : A small bird on a bar looks down***
	**GT1:** A girl going into a wooden building **GT2:** A little girl climbing into a wooden playhouse **GT3:** A little girl in a pink dress going into a wooden cabin **GT4:** A little girl climbing the stairs to her playhouse **GT5:** A little girl climbing the stairs to her playhouse ***D1 : A child going into a house*** ***D2 : A child in a pink dress going into a house*** ***D3 : A child in pink dress is going into a wooden house***

Fig. 5 Samples of image captions generated by the DWCNN based image captioning method with the five ground truth sentences denoted as GT1, GT2, GT3, GT4 and GT5 respectively. **D1**, **D2** and **D3** denotes the descriptions generated by the three configurations -WCNN+LSTM, WCNN+VPAN+LSTM amd WCNN+VAPN+CSE+LSTM, respectively

Abbreviations

WCNN	Wavelet transform based Convolutional Neural Networks
VAPN	Visual Attention Prediction Network
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
CSE	Contextual Spatial Relation Extractor
CRF	Conditional random field
FRMM	Fusion-based Recurrent Multi-Modal
RFNet	Recurrent Fusion Network
CNN	Convolutional Neural Network
CA	Channel attention
SA	Spatial attention
RCNN	Region CNN
RPN	Region Proposal Network
BLEU@N	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation
B@N	BLEU@N
MT	METEOR
R	ROUGE
CD	CIDEr

Acknowledgements
Not applicable.

Author contributions
RS has performed design, implementation and experiments, and written original version of the final manuscript under the guidance, supervision and support of SK and SS. All authors contributed equally in this work and all authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
1. The Flickr 8k datset is available at https://www.kaggle.com/dibyansudiptiman/flickr-8k. 2. The MS COCO dataset is available at: https://cocodataset.org/. 3. The Flickr30k dataset is available at: https://shannon.cs.illinois.edu/DenotationGraph/.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
The authors have no objection in publishing the manuscript if accepted after review.

Competing interests
The authors declare that they have no competing interests.

Received: 24 February 2022 Accepted: 21 January 2023
Published online: 08 February 2023

References

1. Li S, Kulkarni G, Berg TL, Berg AC, Choi Y. Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011, pp. 220–228
2. Lin D. An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 296–304
3. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
4. Jing Z, Kangkang L, Zhe W. Parallel-fusion lstm with synchronous semantic and visual information for image captioning. J Vis Commun Image Represent. 2021;75(8): 103044.
5. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015
6. Gao L, Wang X, Song J, Liu Y. Fused GRU with semantic-temporal attention for video captioning. Neurocomputing. 2020;395:222–8.
7. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, 2018: pp. 6077–6086.
8. Fu K, Jin J, Cui R, Sha F, Zhang C. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2321–34.
9. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. Every picture tells a story: generating sentences from images. In: Computer Vision – ECCV, 2010, pp. 15–29.
10. Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg AC, Berg TL. Baby talk: understanding and generating simple image descriptions. In: CVPR, 2011; pp. 1601–1608.
11. Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé H. Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012; pp. 747–756.
12. Ushiku Y, Harada T, Kuniyoshi Y. Efficient image annotation for automatic sentence generation. In: Proceedings of the 20th ACM International Conference on Multimedia, 2012; pp. 549–558.
13. Mason R, Charniak E. Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2014; pp. 592–598.
14. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: pp. 6298–6306.
15. Guo L, Liu J, Zhu X, Yao P, Lu S, Lu H. Normalized and geometry-aware self-attention network for image captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
16. Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
17. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, 2015; pp. 2048–2057.
18. Yang Z, Yuan Y, Wu Y, Cohen WW, Salakhutdinov RR. Review networks for caption generation. In: Advances in Neural Information Processing Systems, vol. 29, 2016
19. Oruganti RM, Sah S, Pillai S, Ptucha R. Image description through fusion based recurrent multi-modal learning. In: 2016 IEEE International Conference on Image Processing (ICIP), 2016: pp. 3613–3617.
20. Jiang W, Ma L, Jiang Y, Liu W, Zhang T. Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018.
21. Wang, W., Ding, Y., Tian, C.: A novel semantic attribute-based feature for image caption generation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018; pp. 3081–3085
22. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):677–91.

23. Yu L, Zhang J, Wu Q. Dual attention on pyramid feature maps for image captioning. IEEE Transactions on Multimedia; 2021
24. Liu M, Li L, Hu H, Guan W, Tian J. Image caption generation with dual attention mechanism. Inf Process Manag. 2020;57(2): 102178.
25. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):664–76.
26. Deng Z, Jiang Z, Lan R, Huang W, Luo X. Image captioning using densenet network and adaptive attention. Signal Process Image Commun. 2020;85: 115836.
27. Yang M, Zhao W, Xu W, Feng Y, Zhao Z, Chen X, Lei K. Multitask learning for cross-domain image captioning. IEEE Multimedia. 2019;21(4):1047–61.
28. Jiang W, Li X, Hu H, Lu Q, Liu B. Multi-gate attention network for image captioning. IEEE Access. 2021;9:69700–9. https://doi.org/10.1109/ACCESS.2021.3067607.
29. Yang L, Wang H, Tang P, Li Q. Captionnet: a tailor-made recurrent neural network for generating image descriptions. IEEE Trans Multimedia. 2021;23:835–45.
30. Zha Z-J, Liu D, Zhang H, Zhang Y, Wu F. Context-aware visual policy network for fine-grained image captioning. IEEE Trans Pattern Anal Mach Intell. 2022;44(2):710–22.
31. Gao L, Li X, Song J, Shen HT. Hierarchical lstms with adaptive attention for visual captioning. IEEE Trans Pattern Anal Mach Intell. 2020;42(5):1112–31.
32. Yan C, Hao Y, Li L, Yin J, Liu A, Mao Z, Chen Z, Gao X. Task-adaptive attention for image captioning. IEEE Trans Circuits Syst Video Technol. 2022;32(1):43–51.
33. Xiao H, Shi J. Video captioning with adaptive attention and mixed loss optimization. IEEE Access. 2019;7:135757–69.
34. Al-Malla MA, Jafar A, Ghneim N. Image captioning model using attention and object features to mimic human image understanding. J Big Data. 2022;9:20.
35. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y, Gao J. Oscar: object-semantics aligned pre-training for vision-language tasks. In: ECCV; 2020
36. Wang P, Yang A, Men R, Lin J, Bai S, Li Z, Ma J, Zhou C, Zhou J, Yang H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. CoRR; 2022. **abs/2202.03052**
37. Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020.
38. Hu X, Gan Z, Wang J, Yang Z, Liu Z, Lu Y, Wang L. Scaling up vision-language pre-training for image captioning. CoRR 2021. **abs/2111.12233**. arXiv2111.12233
39. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J. Vinvl: Making visual representations matter in vision-language models. CoRR 2021. **abs/2101.00529**
40. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
41. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1. NIPS'15, 2015: 91–99.
42. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; pp. 1179–1195.
43. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res. 2013;47:853–99.
44. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist. 2014;2:67–78.
45. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Doll P, Zitnick CL. Microsoft coco: common objects in context. In: Computer Vision – ECCV 2014; pp. 740–755.
46. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015; pp. 3128–3137.
47. Papineni K, Roukos S, Ward T, Zhu W. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002; pp. 311–318.
48. Lavie A, Agarwal A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, 2007; pp. 228–231.
49. Lin C. ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out, 2004; pp. 74–81.
50. Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based Image Description Evaluation 2015. arXiv1411.5726
51. Sweldens W. The lifting scheme: a custom-design construction of biorthogonal wavelets. Appl Comput Harmon Anal. 1996;3(2):186–200.
52. Deng J, Dong W, Socher R, Li L, L, K, F, L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009; pp. 248–255.
53. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization (2014). arXiv1412.6980
54. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), 2014; pp. 1532–1543. http://www.aclweb.org/anthology/D14-1162
55. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 4651–4659.
56. Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 4904–4912.
57. Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: ECCV 2018; 2017
58. Song H, Zhu J, Jiang Y. avtmnet: adaptive visual-text merging network for image captioning. Comput Electr Eng. 2020;84: 106630.
59. Wang H, Wang H, Xu K. Evolutionary recurrent neural network for image captioning. Neurocomputing. 2020;401:249–56.

60. Hossain MZ, Sohel F, Shiratuddin MF, Laga H, Bennamoun M. Bi-san-cap: Bi-directional self-attention for image captioning. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), 2019; pp. 1–7.
61. Wu J, Chen T, Wu H, Yang Z, Luo G, Lin L. Fine-grained image captioning with global-local discriminative objective. IEEE Trans Multimedia. 2021;23:2413–27.
62. Wang S, Meng Y, Gu Y, Zhang L, Ye X, Tian J, Jiao L. Cascade attention fusion for fine-grained image captioning based on multi-layer lstm. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021; pp. 2245–2249.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com