Attentive Contextual Network for Image Captioning

Jeripothula Prudviraj, Chalavadi Vishnu, and C. Krishna Mohan
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Sangareddy, Hyderabad
{cs17resch01005, cs16m18p000001, ckm }@iith.ac.in

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9533970

***Abstract*—Existing image captioning approaches fail to generate fine-grained captions due to the lack of rich encoding representation of an image. In this paper, we present an attentive contextual network (ACN) to learn the spatially transformed image features and dense multi-scale contextual information of an image to generate semantically meaningful captions. At first, we construct deformable network on intermediate layers of convolutional neural network (CNN) to cultivate spatial invariant features. And the multi-scale contextual features are produced by employing contextual network on top of last layers of CNN. Then, we exploit attention mechanism on contextual network to extract dense contextual features. Further, the extracted spatial and contextual features are combined to encode the holistic representation of an image. Finally, a multi-stage caption decoder with visual attention module is incorporated to generate fine-grained captions. The performance of the proposed approach is demonstrated on COCO dataset, the largest dataset for image captioning.**

***Index Terms*—Image captioning, deformable network, contextual network, attention mechanism, multi-stage LSTM.**

I. Introduction

Automatically describing the content of an image, often termed as image captioning, is one of the primary goals of scene understanding. In recent years, it has received an enormous interest by bringing computer vison (CV) and natural language processing (NLP) together. Image captioning is a challenging task that goes beyond the conventional tasks such as image classification [15] and object detection [6] as it needs to captivate the holistic representation of an image in order to generate fine-grained caption. To achieve holistic representation of an image, we first need to capture the fine-grained visual information and inherent semantic representation of an image. Further, the fine-grained captions can be generated by understanding the semantic relationships among the objects and attending to the prominent image regions. Captioning an image is emerged as a notable research problem due to its wide range of applications in the field of computer vision. It can help visually impaired users, strengthen robotic vision, reinforce the content search, and make it easy to organize & accessible for unstructured visual data.

The task of image captioning learns a probabilistic model over the caption, conditioned on either visual features or combined visual & attribute features of an image. The predominant approach for image captioning task is encoder-decoder framework [24], [28], [37], [39], adopted from machine translation [4]. The encoder-decoder framework employs a convolutional neural network as an encoder to extract the visual information

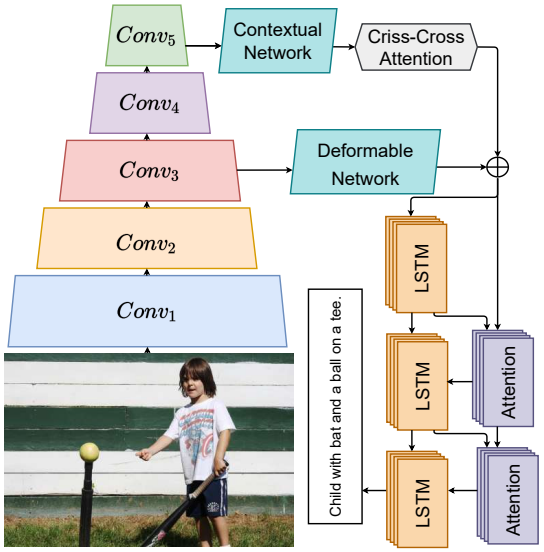

Fig. 1. Overview of the proposed attention-guided approach for image captioning. In brief, we first extract multi-scale contextual features by incorporating dilated convolutions on semantic features. Then, we combine the attention guided spatial and multi-scale contextual features by leveraging position and channel attention mechanisms. The combined features are further fed to the attention based LSTM decoder to generate the caption of an image.

of an image, and the recurrent neural networks are exploited to generate natural language sentence. In order to generate caption of an image, most of the image captioning approaches utilize the semantic representation of an image that is captured from the last convolutional layer of backbone CNN network [15], [19], or the visual object region features extracted from the object detection frameworks [27]. However, these approaches are failed to generate the fine-grained captions of an image due to the lack of holistic representation of an image. The potential drawbacks with the existing approaches are: i) Lack of spatial and contextual information of an image. ii) Fail to incorporate spatially transformed and dense multi-scale contextual features. iii) Utilization of one-stage caption decoder which is hard to generate rich fine-grained captions.

Recent approaches [20], [25] on visualizing the characteristics of each CNN layer demonstrate that the features of early layers contain rich spatial features but lack semantic information [32]. Whereas, the last layers of CNN contribute rich semantic information but fail to incorporate spatial features of small objects [40]. Therefore, we need to capture spatial features along with semantic information in order to obtain

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

fine-grain and coarse-grain details of visual objects. Although this feature fusion approach collects both spatial and semantic details, it is unable to incorporate spatial invariant features and dense scene contextual information. Motivated by the above observations, we propose an attentive contextual network (ACN) to capture spatially transformed image features and dense-contextual information of an image for image captioning task.

The overview of the proposed attentive contextual network (ACN) is shown in Figure 1. The proposed ACN extracts the spatial features from early convolutional layers and semantic information from last layers of CNN. But, the early layers of CNN are sensitive to the geometric transformations such as rotation, pose, object scale, and deformation. These geometric transformations cannot be handled using conventional CNN networks due to the fixed geometric structure of convolutional kernel and max-pool operations. Hence, we incorporate a deformable network [16] on early CNN layers to strengthen transformation modelling capability. Deformable networks are simple and dynamic models that handle both geometric transformations and unknown transformations. Also, it learns the robust representation of objects with different scales and deformations.

Even though, the last layers of CNN contain rich semantic information, they do not preserve multi-scale features and object boundaries information due to the repeated pooling and striding operations in the network [45]. So, we introduce contextual network on top of last convolutional layer of CNN to encode multi-scale contextual information of an image. The contextual network captures the scene contextual information of an image by probing the parallel dilated convolutions at various rates and multiple field-of-views. Further, we leverage an attention mechanism [3] to generate dense contextual features. Finally, the extracted spatial invariant and dense multi-scale contextual features are concatenated and fed to the caption decoder module. The caption decoder module exploits coarse-to-fine multi-stage network to generate refined descriptions of an image. The main contributions of our work are summarized as follows:

- We present attentive contextual network to effectively encode the guided contextual information of an image for caption generation.
- Our encoder-decoder framework incorporates spatial invariant and dense multi-scale contextual features by exploiting deformable and contextual networks.
- The proposed network outperforms the significant works of image captioning on COCO dataset and generates semantically meaningful captions by learning rich holistic representation of an image.

II. Related work

In this section, we first review the several feature encoding approaches that explored spatial & semantic features, spatial invariant features, and multi-scale contextual information for various computer vision tasks. Then, we present the significant works of image captioning.

A. Visual feature encoding approaches

Ma *et al.* [29] exploited rich hierarchical features of deep CNN to boost the accuracy and robustness of visual tracking. These hierarchical features interpret the image pyramid representation and encode the appearance of target objects at multiple levels of abstraction. The combined coarse and fine semantic information via shortcut fusion method is explored in [40] to achieve spatially aware visual details of an image. Zhang *et al.* [32] proposed ExFuse to integrate semantic details into low-level features and high spatial resolution information into high-level features more effectively for segmentation task.

To incorporate the ability to learn the spatial invariant features, Jaderberg *et al.* [46] proposed spatial transformer networks (STNs). These learnable modules explicitly allow the spatial manipulation of data and remove spatial transformations such as affine or perspective without any additional supervision. Inspired by STNs, Dai *et al.* [16] introduced deformable networks to further enhance the transformation modelling capability. The deformable networks generate dense predictions when compared to STNs and further increase the performance over complex vision tasks. Further, various works [11], [46] are proposed to handle deformations and multi-scale variations caused by geometric transformations.

Yu and Koltun [30] introduced dilated convolutions in CNNs to incorporate multi-scale contextual information of an image. Dilated convolutions enlarge the receptive field size exponentially and equip rich contextual information. Si *et al.* [36] proposed multiply spatial fusion network (MSFNet) to extract spatial information and increase the size of receptive field. The spatial pyramid pooling module is introduced in [45] to incorporate multi-scale contextual information by probing parallel dilated convolutions with multiple rates and multiple field-of-views.

B. Image captioning

The classical encoder-decoder framework for image captioning task is explored in [24], where convolutional neural network is employed as encoder and recurrent neural network is served as decoder. Xu *et al.* [34] exploited attention based encoder-decoder framework for image captioning to selectively focus on prominent regions of visual information to generate caption of an image. Chen *et al.* [39] incorporated spatial and channel wise attention in CNN to encode attentive visual information. Further, the word-level and sentence-level attention mechanism is investigated in [43] to generate human like captions. To generate rich fine-grained captions of an image, Gu *et al.* [12] proposed coarse-to-fine multi stage caption generation network with multiple LSTM decoders. Recently, Zhou *et al.* [37] proposed a multi-level visual fusion network (MVF) to interpret visual features as visual knowledge and generate function words for refined captions. The transformer based architecture is utilized in [5] to enrich the visual encoding and caption generation steps.

In contrast to the existing approaches, our proposed method encodes the visual image by incorporating spatial invariant and

dense multi-scale contextual features. And, we leverage multi-stage caption decoder module to generate precise and diverse captions.

III. Proposed approach

In this section, we first describe the classical encoder-decoder framework for caption generation, where the encoder employs convolutional neural network (CNN) to encode the visual representation of an image and decoder leverages recurrent neural network (RNN) based approaches to decode the visual features into sequence of words. Then, we present the proposed attentive contextual network for image captioning task.

A. Encoder-Decoder model for caption generation

Given an image I and its corresponding caption $S = \{w_1, w_2, \ldots, w_N\}$ consists of N words, the encoder-decoder model directly maximizes the objective as

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p\left(S|I;\theta\right), \tag{1}$$

where θ are the parameters of the model. Then, the log likelihood of joint probabilities over all words is defined using chain rule as

$$\log p(S|I) = \sum_{t=1}^{T} \log p(w_t|w_1, \ldots, w_{t-1}, I), \tag{2}$$

here the model parameters are dropped for convenience. Further, each conditional probability is modeled by the RNN based encoder-decoder framework as

$$\log p(w_t|w_1, \ldots, w_{t-1}, I) = f(h_t, c_t), \tag{3}$$

where f is a nonlinear output function which generates the probability of each predicted word w_t. The h_t and c_t are hidden state and context vector of RNN at time t. In this work, we leverage long-short term memory (LSTM) network from the family of RNN to generate caption of an image. The hidden state h_t in LSTM network is modeled as

$$(h_t) = \mathrm{LSTM}(x_t, h_{t-1}, c_{t-1}), \tag{4}$$

where x_t is the input vector. Usually, c_t provides a visual evidence for generating caption of an image and there are two ways to model the context vector i.e., vanilla encoder-decoder and attention based encoder-decoder frameworks.

In a nutshell, the vanilla encoder-decoder framework extracts context vector from the last fully connected layer of convolutional network. This context vector will be same throughout caption generation process and does not depend on the information captured by the RNN decoder module. Whereas, the context vector of attention based encoder-decoder framework depends on both visual encoder and caption decoder modules. And, it selectively focus on prominent regions of an image at each time step of RNN hidden state. In this work, we adopt the attention based encoder-decoder framework to generate guided contextual features for image captioning task.

B. Attentive contextual network for image captioning

In this work, we present an attentive contextual network (ACN) to encode spatially transformed features and multi-scale contextual information of an input image for caption generation task. The proposed attentive contextual network (Figure 2) adopts attention based encoder-decoder framework, where the visual encoder is comprised of deformable network (DN), contextual network (CN), and recurrent criss-cross attention mechanism (RCCAM). And, the caption decoder is constituted with Long-short term memory (LSTM) network. In particular, we incorporate deformable network on spatial features extracted from early layers of backbone network to handle the intra-class variability caused by spatial transformations. And, the contextual network is employed on top of semantic features extracted from the top layer of backbone network to incorporate multi-scale contextual information. Then, a criss-cross attention mechanism [3] is employed on contextual network to attend prominent regions of visual scene information. Further, we fuse spatially transformed features and contextual features to achieve holistic representation of an image. Finally, the multi-stage LSTM network with attention mechanism [12] is used to generate words which are guided by attentive contextual features. In the following subsections, we will elaborate the visual encoder and caption decoder modules.

1) Visual encoder: The visual encoder is a feature encoding network that encodes spatially transformed features and multi-scale scene contextual information of an image by processing input image through various network components. Mainly, the proposed visual encoding framework has five components, namely, backbone network, deformable network, contextual network, recurrent criss-cross attention mechanism, and feature fusion.

a) Backbone network: We utilize pre-trained ResNet [15] as a backbone feature extraction network to extract the visual representation of an image. It has five layers, namely, $Conv_1$, $Conv_2$, $Conv_3$, $Conv_4$, and $Conv_5$, where each $Conv$ layer composed with varied number of Bottle-Neck layers. Usually, the early layers of backbone network hold rich spatial features of small objects but fail to incorporate semantic information of an image. Whereas, the top layers contain the potential semantic information but lack spatial resolutions of objects. To mitigate this issue, we propose a novel feature fusion network that takes advantage of both spatial and semantic features of an image in order to generate a natural language sentence. The proposed framework for image captioning task is illustrated in Figure 2. As shown in Figure, we first remove the fully connected layers of ResNet backbone network and utilize the spatial and semantic features of $Conv$ layers. In particular, we extract the spatial features from $Conv_3$ layer of ResNet backbone network and incorporate a spatial resolution of 32×32 with 512 channels. However, the early layers of backbone network fail to handle the geometric variations such as scaling, rotation, translation, and so on. To address this problem, we introduce deformable network on top $Conv_3$ layer to extract spatially transformed features. Further, we

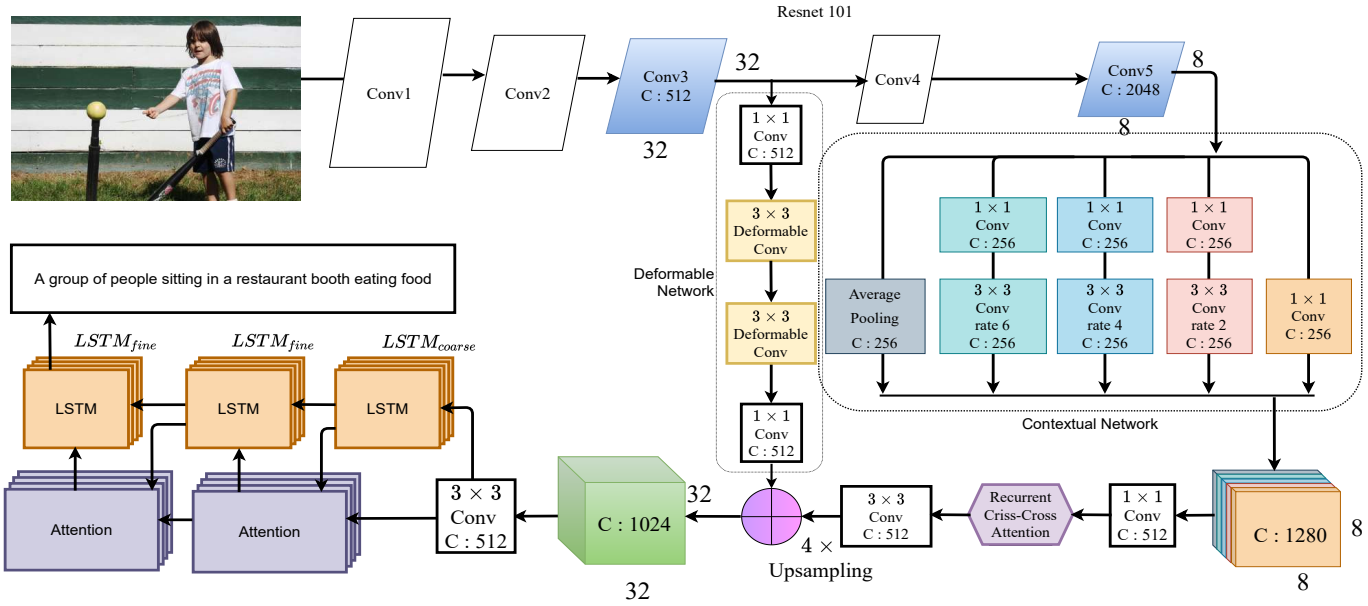

Fig. 2. Framework of the proposed attentive contextual network.

achieve semantic features from $Conv_5$ layer of ResNet with the resolution of 8×8 and 2048 channels. In addition, we employ contextual network on top of $Conv_5$ layer to incorporate multi-scale contextual information of an image. The details of deformable network and contextual network are presented in the following sub-sections.

b) Deformable network: The convolutional networks are widely explored models for image captioning, but their performance is limited by the lack of ability to handle the geometric transformations. Usually, the CNN networks learn to adopt spatial invariance by utilization of large models, data augmentation techniques, and hand-crafted modules such as max-pooling or scale invariant feature transforms (SIFT). However, these hand crafted features learn fixed and known transformations, and fail to process unknown geometric transformations [16]. Further, the large models and hand-crafted design algorithms are difficult, complex, infeasible, and require expensive training for overly complex transformations. In addition, the fixed kernel and max-pool receptive field sizes produce the same activation units by leaving out the fact that the different locations may correspond to objects with different scales and deformations [16].

To address these issues, we employ deformable convolutions on $Conv_3$ of backbone network as they are not truly invariant to large transformations of the input data when compared to the deep convolutional features ($Conv_5$). In the standard convolutions, the deformable convolution module adds 2D offsets at the regular grid sampling locations. These offsets are learned from the preceding convolutional layers and condition the deformation on the input features. The deformable convolutional module can readily place in between convolutional layers and trained end-to-end by simple back-propagation. In our work, we employ two deformable convolutional modules in between two standard 1×1 convolutions.

Given input feature map k, the standard convolution generates the output feature map l using a regular grid $\mathcal{G}$ over the input feature map as

$$l(q_0) = \sum_{q_n \in \mathcal{G}} w(q_n) \cdot k(q_0 + q_n), \tag{5}$$

where w denotes the weight values, q_0 is the location on l, and q_n enumerates on sampling regular grid ($\mathcal{G}$).

Whereas, the deformable convolutions augment the offsets $\Delta\ q_m \mid m = 1, \ldots, M$, where $M = \mod \mathcal{G}$. The Equation 5 is updated for deformable convolution as

$$l(q_0) = \sum_{q_n \in \mathcal{G}} w(q_n) \cdot k(q_0 + q_n + \Delta q_n). \tag{6}$$

In deformable convolutions, the sampling operates on irregular and offset locations ($q_n + \Delta q_n$). Since the offset Δp_n is usually fractional, we use bilinear interpolation in Equation 6 as

$$k(p) = \sum_{s} \mathcal{B}(s, p) \cdot k(s), \tag{7}$$

where q denotes arbitrary location ($k(q_0 + q_n + \Delta q_n)$ for Equation 6), s enumerates all spatial locations in k, and $\mathcal{B}(\cdot, \cdot)$ is the bilinear interpolation kernel.

c) Multi-scale contextual network: Usually, most of the image captioning works [24], [34], [39], [43] generate the caption by utilizing the semantic information of an image that is extracted from top convolutional layer of backbone network. Although the top layer of CNN contains rich semantic information, the finer details of object boundaries diminish due to the multiple pooling and strided convolutional operations. To address this issue, we employ a contextual network on top of $Conv_5$ layer of backbone network. In particular, the contextual

network employs several parallel dilated convolutions [30] on output feature map of $Conv_5$ with different scales in order to capture scene contextual information at multiple scales. The contextual network helps to enlarge the filter's field of view and allows us to control receptive field size of input feature maps. Thus it employs the larger contextual information of an image to the network without increasing the computation time and learnable parameters. Also, the parallel dilated convolutions help to segment the objects at multiple scales and condition each layer to sample the input receptive fields at multiple rates and multiple field-of-views.

Given a two-dimensional signal, for each location j on the output feature map l and weight kernel matrix w, the dilated convolutions are applied on input feature map k as

$$l[j] = \sum_i k[j + r \cdot i]w[i], \tag{8}$$

where r indicates the dilation rate at which we sample the input feature map. Note that $r = 1$ is the special case of dilated convolutions that indicates the standard convolution. As we change dilation rate, the filter's field-of-view modify adaptively.

d) Recurrent Criss-Cross Attention: Although CNN models are exceptionally powerful class of models, they are internally limited to local receptive fields due to the fixed geometric structure and provide short range contextual information. To address this issue, we have introduced contextual network with dilated convolutions at multiple scales. However, our contextual network collects scene contextual information at multiple scales but lacks dense contextual features [3]. Various works utilized [8], [35] attention mechanism to generate dense contextual information by aggregating the contextual information at each position via generated attention maps. Inspired by this, we feed output of contextual network into criss-cross attention module [3] to generate attention guided contextual features. Criss-cross attention mechanism is a memory friendly mechanism and significantly reduces FLOPs when compared to conventional attention mechanisms [39], [42]. In particular, it replaces the dense non-local attention

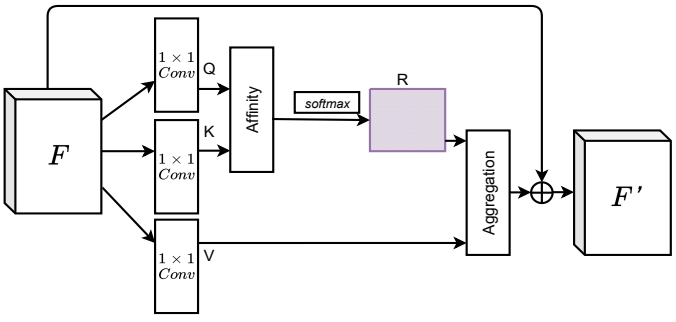

Fig. 3. Criss-cross attention mechanism.

blocks with the sparse attention by sparsely connecting one pixel to other pixels in criss-cross fashion. Further, we take recurrent operation of the criss-cross module to capture the full image dependencies. In a nutshell, the first attention module collects the local context information in horizontal and vertical directions. Then, the recurrent attention module collects the additional information from all other augmented pixels and capture the full image dependencies.

Given an input feature map X, we first obtain a feature map F using a convolutional layer. Then, the feature map F is fed to criss-cross attention module to obtain a new feature map F' that extracts the contextual information at horizontal and vertical direction (criss-cross path). In order to achieve dense contextual feature, we perform the above procedure recurrently and obtain a new feature map F''. The architecture of criss-cross attention mechanism is illustrated in Figure 3. As shown in Figure, given a feature map $F \in \mathbb{R}^{C \times W \times H}$, we generate two feature maps Q and K using 1×1 convolutions, where $\{Q, K\} \in \mathbb{R}^{C' \times W \times H}$. Further, we obtain a feature vector $Q_v \in \mathbb{R}^{C'}$ and $\mho_v \in \mathbb{R}^{(H+W-1) \times C'}$ from each position v in the spatial domain of Q and K, respectively. In order to generate attention map R, we first perform affinity operation as

$$d_{e,v} = Q_v \mho_{e,v}^{\intercal}, \tag{9}$$

where $d_{e,v} \in D$ is the correlation between features Q_v and $\mho_{e,v}$ (e is the e^{th} element of $\mho_v$). Then, we apply softmax layer on D.

In addition, we generate another feature map $V \in \mathbb{R}^{C \times W \times H}$ by employing a convolutional layer on F. Further, we obtain a feature map $\Upsilon_u \in \mathbb{R}^{(H+W-1) \times C}$ by collecting the features in the same row or column at each position v. Then, the aggregation operation collects the all contextual information to generate an output feature vector F' as

$$F'_v = \sum_{e \in |\Upsilon_u|} R_{e,v} \Upsilon_{e,v} + F_u, \tag{10}$$

where F' is the output feature map that contains rich contextual information, Further, we repeat the same criss-cross attention mechanism with the F' in order to obtain dense contextual information (F'').

e) Feature fusion: The purpose of the feature fusion process is to achieve the holistic representation of an image by combining the extracted spatially transformed image features and dense contextual information of an image. As discussed in the previous sections, we extract spatially transformed image features by employing the deformable network on top $Conv_3$ of ResNet backbone network and dense contextual features are achieved by feeding $Conv_5$ features of backbone network to contextual network and recurrent criss-cross attention. In order to fuse these features, we first upsample the contextual features to spatial features using bi-linear interpolation technique. Then, we concatenate both features by stacking up one another and perform 3×3 convolutions before feeding it to caption decoding module.

2) Caption decoding module: Usually, most of the image captioning frameworks use one-stage caption decoder module to generate caption of an image. However, they fail to generate rich fine-grained descriptions due to the lack of intermediate

supervision. To mitigate this issue and effective use of our visual encoder representation, we use a coarse-to-fine multi-stage caption decoder framework [12] for caption generation. Typically, it is equipped with multiple decoders where each decoder component operates on the output of previous stage and produces refined image descriptions progressively. In addition, it addresses the problem of vanishing gradient raised due to multi-stage models and reinforces the intermediate supervision through reinforcement learning. The framework of the coarse-to-fine multi-stage caption decoding module is illustrated in Figure 2. As shown in Figure, it has three stacked long-short term memory (LSTM) networks with attention modules. In particular, the first stage LSTM decoder generates coarse-grained image descriptions, and the successive LSTM decoding network produces the fine-grained descriptions. At each stage of the model, we input attention weights and previous stage hidden vectors to generate more refined captions.

On achieving encoded information of an image using visual encoder, we first learn the coarse decoder LSTM network ($LSTM_c$). At each time step t, the $LSTM_c$ takes the information of previous words, encoded visual representation of an image, and the previous hidden states of LSTM network to generate the caption as

$$c_t^0, h_t^0 = LSTM_c(h_{t-1}^0, i_t^0, w_{t-1}),$$
$$i_t^0 = [f(Z); h_{t-1}^{N_f}], \tag{11}$$

where h_{t-1}^0 and $h_{t-1}^{N_f}$ are the hidden states, c_t^0 is the cell state, w_{t-1} is the previous word, i denotes the decoder stage ($i = 0$ for $LSTM_c$ and $i >= 1$ for fine decoders ($LSTM_f$)), N_f indicates the total number of fine stages, and $f(Z)$ mean pool of visual encoder features. Further, we refine captions using fine stage decoders using attention weights α_t^{i-1}, previous words, and visual information as

$$c_t^i, h_t^i = LSTM_f(h_{t-1}^i, i_t^i, w_{t-1}),$$
$$i_t^i = [g(Z, \alpha_t^{i-1}, h_t^{i-1}, h_t^{i-1}], \tag{12}$$

where, $g(.)$ denotes the function of spatial attention that generates attention guided visual information. On achieving attentive features, the coarse-to-fine LSTM network generates the fine-grained description of an image.

IV. Experimental results

In this section, we verify the efficacy of the proposed attentive contextual network (ACN) using quantitative and qualitative analysis on COCO dataset, the largest dataset of image captioning task. At first, the quantitative analysis is presented by comparing the proposed ACN with the state-of-the art approaches using conventional metrics like BLEU-n, METEOR, ROUGE L, CIDEr-D, and SPICE. Then, we illustrate the generated caption and attention maps of an example image to analyse the performance of the proposed attentive contextual network.

A. Dataset

We conduct experiments on COCO captioning dataset to validate the performance of the proposed attentive contextual network. In our experiment, we use widely adopted COCO data split, which is $113K$ images for training, and $5K$ images for validation & test sets. Further, we evaluate generated captions using standard evaluation metrics like BLEU-n [7] (B-1, B-4), METEOR (MR) [13], ROUGE_L (RL) [21], CIDEr (Cr) [31], SPICE [38]. In a brief, all these metrics take generated caption & reference captions of an input image and evaluate the coherence between $n-$gram word occurrence across the captions.

B. Implementation details

We mainly implemented the proposed attentive contextual network (ACN) using Pytorch [23] framework. In the proposed ACN model, the contextual feature embedding, attention layer embedding, and LSTM hidden and context vectors are fixed to 512 dimension. Further, the ADAM optimizer [26] is used with learning rate of 0.0001 for visual encoder and 0.0003 for caption decoder. Throughout the network, we set batch size to 32 and learnt until the accuracy of the model does not change on validation set for 15 epochs. Finally, we employ decay rate when the model does not improve for 6 epochs.

For our visual encoder, we use ResNet-101 [15] network as backbone network which is pre-trained on Imagenet [33]. From ResNet backbone network, we first extract the spatial ($Conv_3$) and semantic $Conv_5$ features of size $32\times32\times512$ and $8 \times 8 \times 2048$, respectively. On extracting spatial features, we employ deformable network with two standard and deformable convolutions. We use 1×1 filters with 512 channels for standard convolutions. And, the deformable convolutions are employed using 3×3 filters. Further, we incorporate contextual network with several parallel dilated convolutions on $Conv_5$ feature of ResNet backbone network. At first, we employ 1×1 convolutions and 3×3 convolutions with dilation rates of $2, 4$, and 6. Along with dilated convolutions, we add one convolutional layer of 1×1 filter and average pooling to incorporate features from various field-of-views. The convolutional layers of entire contextual network are set to 256 channels and concatenated all 5 layers before feeding it to criss-cross attention module. The number of channels are reduced to 512 before feeding contextual feature to criss-cross attention module and maintained the same number of channels in feature fusion process.

C. Quantitative results

In this work, we present the attentive contextual network (ACN) for image captioning task that incorporates the spatially transformed and multi-scale contextual features of an image to generate natural language sentence. The proposed ACN model generates the spatially invariant and semantically rich features. Table I presents the performance comparison of the proposed ACN model with the state-of-the-art methods of image captioning using standard evaluation metrics. From the Table, we can observe that the proposed ACN approach

is outperforming the previous state-of-the-art approaches in terms of all metrics due to incorporation of rich visual and textual information. In particular, we compare the performance

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH WITH THE STATE-OF-THE-ART METHODS ON THE COCO DATASET, WHERE B-N, MR, RL, CR, AND SP ARE SHORT FOR BLEU-N, METEOR, ROUGE L, CIDER-D, AND SPICE, RESPECTIVELY

	B-1	B-4	MR	RL	Cr	SP
NIC [24]	-	27.7	23.7	-	85.5	-
SCA-CNN [39]	71.9	31.1	25.0	-	-	-
DAIC(RL) [43]	77.6	35.4	26.7	56.4	116.5	-
Up-Down [10]	77.2	36.2	27.0	56.4	113.5	20.3
VRAtt [14]	80.1	37.2	27.9	61.3	121.8	21.9
VREA [28]	80.2	37.4	28.1	57.2	122.1	21.9
MVF [37]	80.5	38.5	28.2	58.1	128.1	22.1
AOA [2]	80.2	38.9	29.2	58.8	129.8	22.4
M^2T [5]	80.8	39.1	29.2	58.6	131.2	22.6
Ours-ACN	**81.2**	**40.1**	**29.6**	**59.2**	**133.3**	**23.1**

of the proposed ACN with the recent encoder-decoder models like double attention image captioning model (DAIC) [43], visual relationship attention (VRAtt) [14], visual relational reasoning (VREA) [28], multi-level visual fusion (MVF) [37], and meshed-memory transformer models (M^2T) [5]. Here, the DAIC [43] model investigates the word level attention and VRAtt [14] utilizes the visual relationship attention for image captioning. Further, multi-level visual features are explored in MVF [37] for caption generation. Finally, the M^2T [5] reported the state-of-the-art performance using transformers in caption decoder module. In contrast to other works, we utilize spatially transformed image features and multi-scale contextual information to encode the visual representation of an image. Further, the multi-stage caption decoder is used to generate semantically meaningful captions.

D. Qualitative results

In this section, we present the qualitative analysis of the proposed attentive contextual network (ACN) through generated caption and attention maps of each generated word for given input image. The attention maps and generated words are shown in Figure 4. From the Figure, we can observe that the proposed approach is attending to relevant image regions of each generated word. Also, we can infer that the generated caption is precise, diverse, and semantically meaningful. Specifically, the generated words "adult", "baby", "dirt", "field", and "elephant" demonstrate that the proposed approach is able to generate fine-grained words by utilizing the holistic representation of an image.

Generated caption: an adult elephant and a baby elephant standing in a dirt field.

Fig. 4. Illustration of attention maps and generated captions of test image.

V. CONCLUSION

Most of the image captioning approaches utilize either attention guided image/object level features [10], [14], or visual attributes [1], [37] with semantic information for describing the gist of an image. However, these methods fail to incorporate spatial information of small objects and multi-scale contextual information of image. To address this issue, we propose a novel attentive contextual network (ACN) for image captioning. The proposed ACN approach incorporates attention guided dense visual features for caption generation task. In particular, we first extract spatial and semantic features from backbone network. Then, we refine the extracted features by employing deformable network and contextual network. Further, we utilize coarse-to-fine multi-stage caption decoder to generate fine grained captions. Finally, we demonstrate the efficacy of the proposed approach on COCO dataset. Especially, the deformable network achieves the spatially transformed features and contextual network provides the dense multi-scale contextual information of an image. The stack of LSTM networks in multi-stage caption decoder module incorporates intermediate supervision and handle the vanishing gradient problem.

REFERENCES

[1] Gan, Zhe, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng., "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 5630-5639, 2017.

[2] Huang, Lun, Wenmin Wang, Jie Chen, and Xiao-Yong Wei., "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4634-4643. 2019.

[3] Huang, Zilong, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu., "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603-612. 2019.

[4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio., "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[5] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara., "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , pp. 10578-10587. 2020.

[6] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie., "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125, 2017.

[7] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu., "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics,* pp. 311-318, 2002.

[8] Fu, Jun, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146-3154. 2019.

[9] Song, Jifei, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales., "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5551-5560, 2017.

[10] Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang., "Bottom-up and top-down attention for image captioning and visual question answering." in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077-6086. 2018.

[11] Lu, Yu, Muyan Feng, Ming Wu, and Chuang Zhang., "C-DLinkNet: considering multi-level semantic features for human parsing," *arXiv preprint arXiv:2001.11690,* 2020.

[12] Gu, Jiuxiang, Jianfei Cai, Gang Wang, and Tsuhan Chen., "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.

[13] Banerjee, Satanjeev, and Alon Lavie., "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,* pp. 65-72, 2005.

[14] Zhang, Zongjian, Yang Wang, Qiang Wu, and Fang Chen., "Visual Relationship Attention for Image Captioning," in 2019 *International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1-8.

[15] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[16] Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei., "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764-773. 2017.

[17] Zhang, Zhenli, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun., "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV),* pp. 269-284, 2018.

[18] Wei, Haiyang, Zhixin Li, Canlong Zhang, Tao Zhou, and Yu Quan., "Image captioning based on sentence-level and word-level attention," in 2019 *International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1-8.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZbigniewWojna., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 2818–2826, 2016.

[20] Mahendran, Aravindh, and Andrea Vedaldi., "Understanding deep image representations by inverting them." in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 5188-5196, 2015.

[21] Lin, Chin-Yew, and Eduard Hovy., "Manual and automatic evaluation of summaries," in *Proceedings of the ACL-02 Workshop on Automatic Summarization, Association for Computational Linguistics,* Volume 4, pp. 45-51, 2002.

[22] Yu, Fisher, and Vladlen Koltun., "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[23] Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer., "Automatic differentiation in pytorch," 2017.

[24] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan., "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156-3164. 2015.

[25] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge., "Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks," in *Bernstein Conference 2015,* pp. 219-219, 2015.

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

[27] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun., "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[28] Pei, Haolei, Qiaohong Chen, Ji Wang, Qi Sun, and Yubo Jia., "Visual Relational Reasoning for Image Caption," in 2020 *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8.

[29] Ma, Chao, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang., "Robust visual tracking via hierarchical convolutional features," *arXiv preprint arXiv:1707.03816,* 2017.

[30] Yu, Fisher, and Vladlen Koltun., "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[31] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh., "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 4566-4575, 2015.

[32] Zhang, Zhenli, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun., "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV),* pp. 269-284, 2018.

[33] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton., "Imagenet classification with deep convolutional neural networks," *Communications of the ACM 60,* vol no. 6, pp. 84-90, 2017.

[34] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio., "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048-2057. PMLR, 2015.

[35] Zhao, Hengshuang, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia., "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 267-283. 2018.

[36] Si, Haiyang, Zhiqiang Zhang, Feifan Lv, Gang Yu, and Feng Lu., "Real-Time Semantic Segmentation via Multiply Spatial Fusion Network," *arXiv preprint arXiv:1911.07217,* 2019.

[37] Zhou, Dongming, Canlong Zhang, Zhixin Li, and Zhiwen Wang., "Multi-level Visual Fusion Networks for Image Captioning," in 2020 *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8.

[38] Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould., "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*, pp. 382-398. Springer, Cham, 2016.

[39] Chen, Long, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua., "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659-5667. 2017.

[40] Song, Jifei, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales., "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision,* 2017, pp. 5551-5560.

[41] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV),* pp. 801-818, 2018.

[42] Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He., "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794-7803. 2018.

[43] Wei, Haiyang, Zhixin Li, Canlong Zhang, Tao Zhou, and Yu Quan., "Image captioning based on sentence-level and word-level attention," in 2019 *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2019.

[44] Annunziata, Roberto, Christos Sagonas, and Jacques Calì., "DeST-Net: Densely fused spatial transformer networks," *arXiv preprint arXiv:1807.04050,* 2018.

[45] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV),* pp. 801-818, 2018.

[46] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman., "Spatial transformer networks," In *Advances in neural information processing systems,* pp. 2017-2025, 2015.