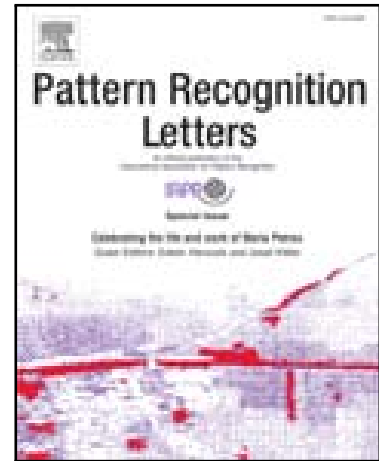


## Accepted Manuscript

A Hierarchical and Regional Deep Learning Architecture for Image Description Generation

Philip Kinghorn , Li Zhang , Ling Shao

PII: S0167-8655(17)30324-0  
DOI: [10.1016/j.patrec.2017.09.013](https://doi.org/10.1016/j.patrec.2017.09.013)  
Reference: PATREC 6925



To appear in: *Pattern Recognition Letters*

Received date: 9 March 2017  
Revised date: 5 July 2017  
Accepted date: 6 September 2017

Please cite this article as: Philip Kinghorn , Li Zhang , Ling Shao , A Hierarchical and Regional Deep Learning Architecture for Image Description Generation, *Pattern Recognition Letters* (2017), doi: [10.1016/j.patrec.2017.09.013](https://doi.org/10.1016/j.patrec.2017.09.013)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

## 9 Pattern Recognition Letters Authorship Confirmation

Please save a copy of this file, complete and upload as the “Confirmation of Authorship” file.

As corresponding author I, Li Zhang \_\_\_\_\_, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Li Zhang \_\_\_\_\_ Date 25<sup>th</sup> Feb 2017 \_\_\_\_\_

---

List any pre-prints: N/A.

---

**Relevant Conference publication(s) (submitted, accepted, or published):**

This research is newly proposed and has not been reported in any previous conference submissions.

**Justification for re-publication:**

N/A.

**Research Highlights (Required)**

To create your highlights, please type over the instructions in the template box below:

- A two-stage deep network model is proposed for image description generation.
- It performs simultaneous human and object attribute labelling for ROIs.
- It is equipped to translate regional information into detailed image descriptions.

ACCEPTED MANUSCRIPT



Pattern Recognition Letters  
journal homepage: [www.elsevier.com](http://www.elsevier.com)

## A Hierarchical and Regional Deep Learning Architecture for Image Description Generation

Philip Kinghorn<sup>a</sup>, Li Zhang<sup>a,\*</sup> and Ling Shao<sup>b</sup>

<sup>a</sup>*Department of Computer and Information Sciences, Northumbria University, Newcastle, UK, NE1 8ST.*

<sup>b</sup>*School of Computing Sciences, University of East Anglia, Norwich, UK, NR4 7TJ.*

### ABSTRACT

This research proposes a distinctive deep learning network architecture for image captioning and description generation. Specifically, we propose a hierarchically trained deep network in order to increase the fluidity and descriptive nature of the generated image captions. The proposed deep network consists of initial regional proposal generation and two key stages for image description generation. The initial regional proposal generation is based upon the Region Proposal Network from the Faster R-CNN. This process generates regions of interest that are then used to annotate and classify human and object attributes. The first key stage of the proposed system conducts detailed label description generation for each region of interest. The second stage uses a Recurrent Neural Network (RNN)-based encoder-decoder structure to translate these regional descriptions into a full image description. Especially, the proposed deep network model can label scenes, objects, human and object attributes, simultaneously, which is achieved through multiple individually trained RNNs.

The empirical results indicate that our work is comparable to existing research and outperforms state-of-the-art existing methods considerably when evaluated with out-of-domain images from the IAPR TC-12 dataset, especially considering that our system is not trained on images from any of the image captioning datasets. When evaluated with several well-known evaluation metrics, the proposed system achieves an improvement of ~60% at BLEU-1 over existing methods on the IAPR TC-12 dataset. Moreover, compared with related methods, the proposed deep network requires substantially fewer data samples for training, leading to a much-reduced computational cost.

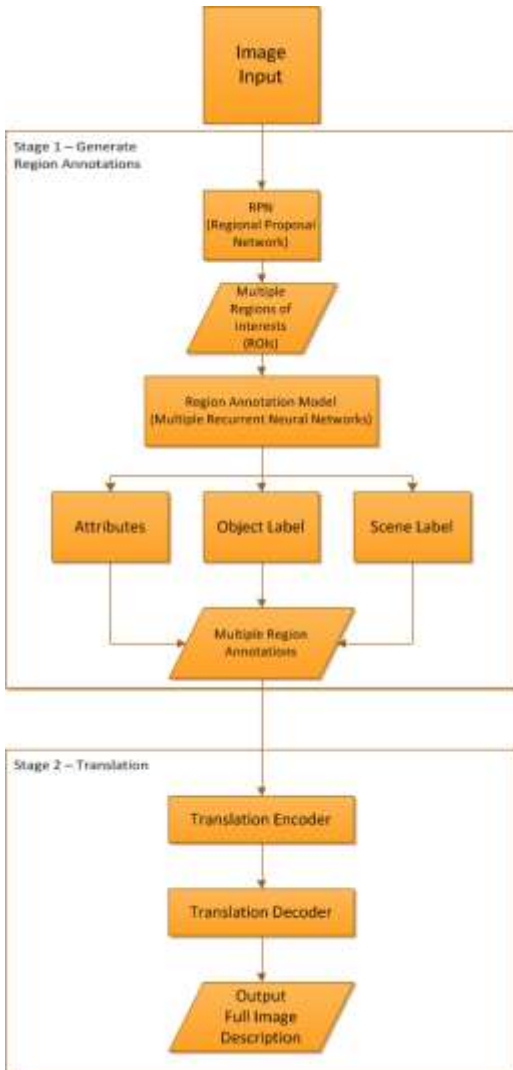
2016 Elsevier Ltd. All rights reserved.

\* Corresponding author. Tel.: +44-191-243-7089; e-mail: [li.zhang@northumbria.ac.uk](mailto:li.zhang@northumbria.ac.uk)

## 1. Introduction

Image captioning is one of the uprising but also challenging research areas for deep learning. A system that can not only accurately label image regions but also scale to whole image description shows great potential in diverse applications such as news or medical image annotation and automatic scripts generation for movies. Many existing research and publicly available datasets were tailored for brief image captioning so far [1, 2]. It is still a challenging task to generate detailed and refined descriptions for both relevant regions and the whole image.

Therefore, in this research, we aim to address the above challenges and propose a novel compact deep network architecture for detailed image description generation. The proposed deep network is composed of multiple Convolutional Neural Networks (CNNs) [3] in combination with Recurrent Neural Networks (RNNs) [4], specifically Long Short Term Memory networks (LSTM) [5] and Gated Recurrent Network (GRU) [6], for image caption generation. It is capable of performing object and scene classification, as well as human and object attribute prediction, simultaneously. Especially, the simultaneous generation of human and object attributes provides rich and detailed descriptions of image regions, and equips the proposed system with impressive capabilities to deal with out-of-domain queries.



**Fig. 1.** The high-level overview of the proposed image description generation system.

The overall architecture of the proposed deep network is presented in Fig. 1. The system is composed of multiple stages (including pre-processing and two key stages), the combination of which allows for the full functionality. The initial processing of the system uses the Region Proposal Network (RPN) [7] to generate multiple regional proposals (i.e. regions of interest), that are likely to contain objects or people. The first key stage of the model conducts object and scene classification and attribute prediction. This stage combines the extracted regional features with word vectors in an RNN for attribute prediction, and utilizes the same regional features for scene and object labelling. The second stage is used for language ‘conversion’. It converts the generated attributes and other class labels into fully descriptive image captions.

In order to enhance the system’s generalization and scalability, instead of trained using existing image caption datasets such as Flickr [8] or MSCOCO [9], the overall proposed deep architecture is hierarchically trained upon multiple different datasets from different domains. These datasets have their individual dedications to be used within a particular domain, e.g. attribute datasets are originally purely dedicated to attribute prediction applications.

Using multiple datasets allows for several benefits. As an example, the proposed system ensures that there is reduced offline training, as compared to other end-to-end and composite methods. Our proposed model ideally would require one dataset that would be annotated with all the functionality of the system. The closest dataset currently is Visual Genome (VG) [10], which is incredibly large with ~110,000 images, however no full image captions are provided, despite all other desired features. Therefore, the proposed system is trained on multiple smaller irrelevant image datasets, yet still provides a competitive outcome to systems that are trained on a single image captioning dataset. Our proposed region-based method allows for increased functionality, including longer image descriptions with regional details, and simultaneous region and full image description generation. Overall, the system shows great diversity for image caption/description generation with more efficient training and testing in comparison to existing methods.

Moreover, as the proposed system is not specifically trained on image-to-caption datasets, such as Flickr [8] or MSCOCO [9], another main advantage of the system is that it can handle out-of-domain images efficiently. This ensures that a reasonable detailed description can be generated for most images passed on to the system regardless of its source to increase the system’s robustness. Finally, the main contributions of our research are summarized as follows:

- A novel deep architecture for image region annotation is proposed. It generates not only regional annotations but also integrates the regional captioning into full image descriptions.
- The proposed deep network has a more efficient training process and shows great robustness and efficacy in dealing with out-of-domain images. It has also been deployed and integrated with the vision API of a humanoid robot to indicate its effectiveness in real-life settings.

## 2. Related Work

Recent research in the image captioning domain has typically been dominated by the use of CNN + RNN [1, 2, 11], which can be split into two general categories, i.e. composite and end-to-end models. The composite method [12] is comparatively simpler,

and utilizes a template system to import the detections into multiple pre-defined structures. It generally produces rigid sounding image captions. However, motivated by recent superior research, and similar to our proposed network model, the second method focuses on end-to-end architectures, which are capable of generating an image caption in one pass of the model. Such systems tend to be built upon multiple components and use the outputs of all individual components to produce the final caption. Such end-to-end systems are fluid and natural compared to the above template-based methods, however require a comparatively higher computational cost.

However, the captions generated by these two existing composite and end-to-end models are also small and uninformative, typically around 10 words per image [9]. In this research, we aim to improve upon this limitation and generate a system to describe images in a much higher degree of details.

### 2.1. Recent Research on Image Captioning

Recent work of Johnson et al. [13] introduced a dense captioning system to this field. Rather than captioning an image, their work captions many individual regions with rich annotations, e.g. objects and attributes. This is achieved with a localization layer which acts as a region proposal generator to annotate image regions. This layer was developed based upon the research of Ren [7], in which a Region Proposal Network was trained to generate the regional proposals rather than relying on the existing less efficient techniques such as EdgeBoxes of Selective Search.

Tan and Chan [14] proposed a system similar to the existing work of Johnson et al. [13]. However, rather than utilizing the typical word based approaches that RNNs tend to adopt, their model encodes the sentence as a combination of both phrases and words for image caption generation.

Recent work of Matsuo et al. [15] showed initial exploration of quantitative natural language descriptions utilizing human brain activities. Owing to the lack of brain activity datasets for deep learning research, the work re-uses frameworks of Vinyals et al. [1] and Xu et al. [11]. Overall, the work synchronized image datasets from movies and brain activity data from an fMRI

scanner, and relied on the MRI data to generate descriptions.

Fang et al. [16] proposed a caption generation system utilizing a bag of words method. Their work implements multiple instance learning and uses visual classifiers for words that commonly appear in existing captions. In addition, their system treats the caption/description generation as an optimization problem. It takes the previously generated words and then finds a sentence/caption that has the highest likelihood to caption the image that contains every word it has detected.

Tran et al. [17] have produced a system that could richly caption images. Their research claims to be able to detect and classify a large range of visual concepts. This includes specific locations, as well as specific persons such as celebrities or people of influence. Their framework consists of a compositional approach. It combines a feature extracting CNN, which passes features to their visual concept network that was trained on 700 visual concepts, such as celebrities and landmarks. It then follows similar research in the field and passes these into a language model. Their work has an advantage, i.e. if the system's confidence is low, instead of generating a rich caption, it can produce a simpler caption that can essentially annotate/list the objects within the image.

Users' vision has also been considered in attempt to improve existing image captioning research. Human gaze has been explored for tasks such as localization in the form of attention [11]. For instance, Sugano and Bulling [18] explored gaze-assisted image captioning by examining the relationship between human gaze and attention mechanisms. Also, the attention mechanism has been adopted by the work of Xu et al. [11].

Image captioning has also been explored recently in a multi-lingual set-up [19], in order to caption and describe images in more than one target languages. This has been explored both as image caption and machine translation, as well as modifying the RNN to generate multiple language outputs.

The deep network proposed in this research is motivated by the above existing frameworks. The individual key aspects of some closely related research frameworks are summarized in

**Table 1.** The methodologies of existing research frameworks

Related work	Methodologies	Contributions
Ren [7]	Region Proposal Network (RPN) – Allowing accurate and near cost free region proposals. Faster and more reliable than methods such as Selective Search.	<ul style="list-style-type: none"> <li>Constructing a RPN, which is a fully connected network trained end-to-end specifically on region proposals.</li> <li>Proposing a deep model that alternates between fine-tuning the RPN and the object detector.</li> </ul>
Xu et al. [11]	Saliency/Attention model – Focusing on a region similar to that in which the human vision behaves	<ul style="list-style-type: none"> <li>Introducing both soft and hard attention mechanisms, as well as showing how ‘where’ and ‘what’ can be used to gain insight and interpret the results from the framework by visualizing where the model was ‘looking’.</li> </ul>
Johnson et al. [13]	Regional description – Applying caption techniques to individual image regions	<ul style="list-style-type: none"> <li>Introducing a dense localization layer that can be implanted into existing CNN models.</li> <li>Introducing a new large-scale dataset (i.e. Visual Genome).</li> </ul>
Tan and Chan [14]	Phrase based LSTM – Encoding sequences of phrases and words	<ul style="list-style-type: none"> <li>Proposing a novel phrase based LSTM in which the image is encoded in three stages, i.e. chunking of the image, phrase composition as a vector representation and encoding the sentence based on the image, words, and phrases.</li> </ul>
Fang et al. [16]	Determining salient content and knowing which image contents are interesting or novel using contextual common sense knowledge.	<ul style="list-style-type: none"> <li>Re-ranking word detectors that capture global semantics.</li> </ul>
Tran et al. [17]	Rich description – Adding specifics to image, such as person and location	<ul style="list-style-type: none"> <li>Presenting a caption model for open domain images, which utilizes a composite approach.</li> <li>Enriching existing frameworks with visual concepts such as landmarks and celebrity identification.</li> </ul>
Sugano and Bulling [18]	Employing gaze annotated image inputs to generate gaze assisted captioning	<ul style="list-style-type: none"> <li>Providing an analysis of the relation between object and scene recognition models and human gaze, as well as presenting a novel gaze assisted attention framework.</li> </ul>

Table 1.

### 3. The Proposed Deep Network for Image Description Generation

This research proposes a hierarchical deep network with the intention to produce image description with a great level of details. It integrates multiple CNNs with particular types of RNNs such as LSTM and GRU, for image description generation. We introduce each key stage of the proposed network below.

#### 3.1. Model Training

Our proposed framework is loosely categorized as an end-to-end system. It presents a unified model that can generate not only descriptive region annotations, as DenseCap [13], but also full image descriptions, as Google NIC [1] and NeuralTalk [2]. Due to the large and complex nature of the proposed model, and the fact that the model is trained on multiple datasets, the proposed system has to be trained hierarchically.

This process first involves freezing multiple sections and branches of the model, before training and fine-tuning the desired weights on the relevant branches with the relevant data. This leads to a large amount of training data being utilized, across multiple datasets, including the generation and use of dummy data for the unaltered branches. The freezing and training of certain layers at the training stage depend upon the dataset currently in use.

VGG [20] is a popular CNN and commonly used for object classification. It has a high top-5 accuracy which means that these weights of its layers can directly be inserted into the

corresponding layers of our model for object classification, as shown in Fig. 2. However, VGG's success can be in part related to the discriminative features it extracts before applying its fully connected layers. We intend to utilize these features for more than just object and scene classification, but also subsequent attribute prediction.

This is achieved as previously stated by unaltering the convolution layers of VGG, so that the extracted features will remain the same. These features are then used to train the scene classification using the fully connected layers on the given dataset. A list of datasets used for the training of each key stage of the proposed model is provided in Section 3.1.1.

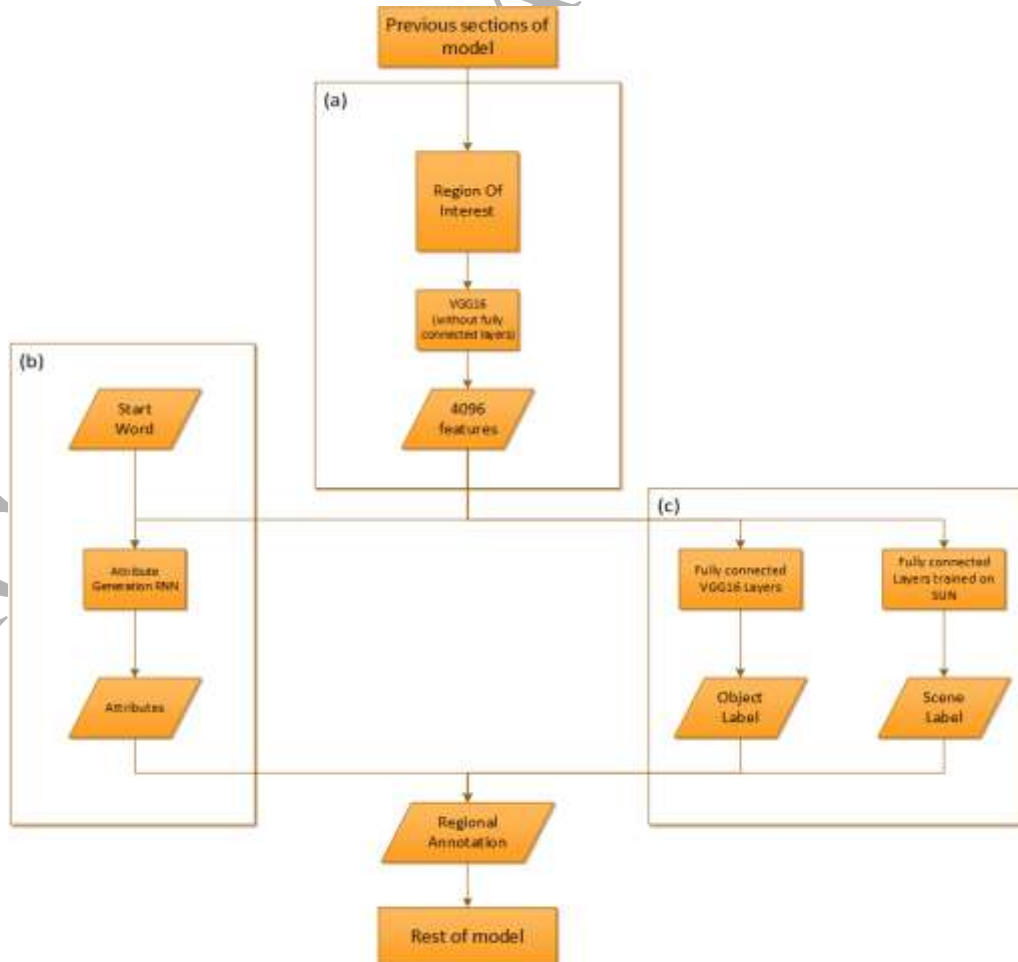
Human and object attributes are also trained in our current model configuration. This stage uses word encoding in combination with the extracted features in order to generate regional attribute labels. Again, all layers which are irrelevant during training are isolated or unaltered.

When training the deep neural networks to obtain a target output, we are only training certain layers, while the frozen layers are given dummy data. These will not update the weights in any way yet still allow the training of the necessary layers to take place. This process is repeated for all relevant branches of the model.

Although the model is currently trained hierarchically, in future work, if a dataset exists that covers all needed outcomes, the model could theoretically be trained in an end-to-end fashion.

#### 3.1.1. Model Datasets

In Fig. 2, the model has been split into a number of components. We highlight the datasets used to train the relevant



**Fig. 2.** (a) The initial process of extracting features from the generated region of interest. (b) The process of combining the features and word vectors to generate region attributes. (c) Using the re-added VGG layers to generate object labels and the trained scene layers to generate scene labels.

layers below.

- ImageNet (objects labels) [21] – This dataset consists of around half a million images, for 200 objects. VGG is also pre-trained on this dataset.
- PubFig [22] – This large facial image dataset initially has 79 attributes for each of the ~60,000 images. We only use a small subset of the available human attributes, although a large subset of the images is used in this research.
- ImageNet (object attributes) [21] – A small portion of 10,000 images from the full ImageNet dataset is paired with 10 object attributes. All of these images are used for training in our work.
- MSCOCO/IAPR TC-12 [9, 23] captions – We collect captions from MSCOCO and IAPR TC-12 for the training of the caption generator in this work.
- SUN scene dataset [24] – This dataset consists of more than 100,000 images of 397 scene categories. We use a subset of the available images, i.e. 10,000 images, for training.

### 3.2. Architecture

The initial processing of the system requires regions to be collected and cropped. These regions are likely to contain objects or people for the system to annotate. This stage involves region proposal generation. In this research, it is implemented by the Faster R-CNN [7]. The RPN within the Faster R-CNN is essentially a powerful neural network that generates bounding box regions and confidence scores. It produces a high score when the system believes the region contains an object or something of interest. The number of regional proposals passed on to the next stage is determined by several factors, such as the size and the complexity of the image as well as the generated confidence measures.

After generating regional proposals, the rest of the proposed model is split into two key stages. The first stage generates detailed regional labels, and the second stage translates these region labels into a full description. We introduce these two stages in detail in the following.

The first key stage accepts two inputs, i.e. a start word and an image. It generates attribute and object labels word by word and the sequence is fed back into the network to produce all attribute and object labels. The second stage is based upon a common machine translation approach to ‘translate’ the labels from source (i.e. region labels) to a full description. Therefore, the final generated description is expected to be more detailed than that of existing research.

These two stages, described for the rest of this work as regional and translation models, can be further broken down into a number of branches that are responsible for a specific task, i.e. attribute prediction and scene and object classification. These branches are explained in detail below.

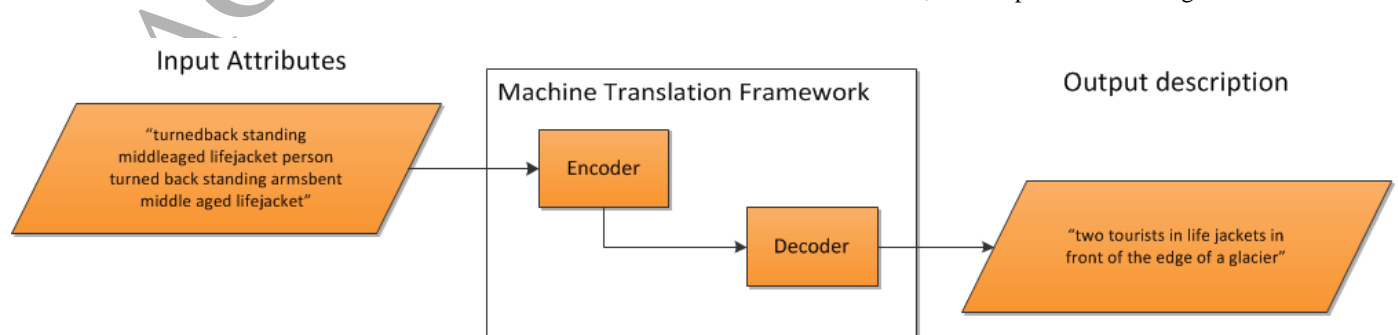
The regional model in the first stage can be split into three branches, as shown in Fig. 2. The first section, as shown in Fig. 2(a), shows the region of interest being passed into the VGG feature extraction network to generate the discriminative features as used in the subsequent stages of the model. The left branch, as indicated in Fig. 2(b), learns word feature vectors, which are ultimately used for attribute generation. This branch converts word integer positions within the vocabulary into a fixed size vector. In our work, this is a 128-dimensional vector.

The vector is passed through a GRU [6] accompanied with fully connected layers to generate an output at each time step. The next stage in this branch involves the image features. The outputs of GRU and sequence image features are then merged and combined before being passed through the subsequent language generating neural network. This network consists of LSTM [5] layers in combination with a fully connected layer in which the next word in the sequence is generated. The aim of this branch is to generate human and object attribute labels of the input regions which are later used in the translation stage.

The first part of the right branch, as illustrated in Fig. 2(c), specifically classifies object labels. This branch uses 4096 image features extracted from the VGG16 CNN. This branch initially has the last two layers removed, which consist of a dropout and a fully connected layer, in order to deliver the features. However, these layers are re-added subsequently in the architecture so that captions and object labels can be simultaneously generated, and later concatenated with the attribute labels to deliver an overall output. On the same branch, the extracted image features are used to classify a scene label. Furthermore, the far right section, in Fig. 2(c), is used to classify the scene and it generally uses the whole image as a region. To achieve classification, the layers of this branch are fine-tuned on the SUN dataset [24]. The initial stage of this process is the same as the object classification as it collects 4096 discriminative features from the VGG16 CNN.

This abovementioned processes rely on multiple training stages. However, the second stage of the model, i.e. the ‘translation’ or ‘conversion’ stage, only requires one training step, while the entire top section, previously discussed, remains as is, unaltered by this stage.

This translation model in the second stage shown in Fig. 3 follows the work of Bahdanu et al. [25] in which a single neural network architecture consists of an improved encoder-decoder. This encoder-decoder architecture is originally designed for machine translation, and outperforms existing statistical machine



**Fig. 3.** An example of the generated input regional attributes and its translated output description



translation approaches such as Koehn [26] and Sutskever et al. [27].

This model structure of encoder-decoder [28] is still used in the typical fashion of encoding the source text into a vector, and

decoding the vector to generate the target text. In this research, the source consists of extracted attribute labels together with scene and image information, with the target being a detailed image description.

### Good Results:



**Ours:** three grown-ups in a flat landscape with a mountain in the background

**NTalk:** a man standing on a beach holding a surfboard

**NIC:** a man and a woman are sitting on a rock overlooking a lake



**Ours:** a man is standing on a black rock with a sandy desert, a green valley and a mountain range and clouds in

**NTalk:** a man sitting on a bench in the middle of a field

**NIC:** a man and a woman are sitting on a rock overlooking a lake



**Ours:** four female tourists are posing with a large, dark brown mountain with a snow covered peak in the background

**NTalk:** a man and a woman are walking on a beach

**NIC:** a man is standing on a snow covered mountain



**Ours:** a grey and light brown house with a small very dense vegetation behind it

**NTalk:** a man and a woman are sitting on a bench in a park

**NIC:** a man sitting on a bench with a dog

### Reasonable Results:



**Ours:** a man in a black jacket with a grey rain jacket on a head on a grey brick on a bridge over mountains in the background

**NTalk:** a man and a woman are standing on a rocky path

**NIC:** a man and a woman standing next to each other



**Ours:** a man green mountain landscape with a few towns with a brown, bald mountain range in the background

**NTalk:** a young girl in a pink dress is walking on a path

**NIC:** a man and a woman standing next to a man

**Fig. 4.** Example outputs generated by the proposed system, Google NIC and NeuralTalk (referred as NTalk) on the IAPR TC-12 dataset

As previously discussed, the encoding structure is used to encode the attribute and class labels into a vector. It initially encodes into a sequence of vectors, of which a subset is adaptively chosen for use during the decoding stage. This is followed by the decoder which uses this subset to generate an image description. This encoder-decoder processing is opposed to a fixed length vector, which is determined to be a bottle neck problem in existing research. An example of generated regional attributes and its associated translated output is shown in Fig. 3. As can be seen in Fig. 3, the generated input attributes describe two middle aged persons in life jackets. This is translated and the system generates ‘two tourists in life jackets in front of the edge of a glacier’, owing to the training of the caption generation system where the glacier information is inferred in the output description. The scene information is omitted in the attributes when a confidence value is not reached. The above two-stage deep network implementation is utilized in this research owing to the stated improvements over other existing work for the generation of sentences that are longer and more descriptive than other works.

### 3.3. The Deployment to the Robot Platform

A real-life application of this system has also been initially explored, by combining the proposed model with the vision SDK of a humanoid robot, NAO NextGen H25. The NAO robot has a powerful CPU processor and camera sensors to allow for real-time image processing and better low light perception. The integration of the proposed system with the robot’s SDK enables the robot to conduct health monitoring, e.g. to identify falling subjects and describe users’ environment to promote personalized human robot interaction. It also enables the evaluation of this proposed system in diverse real-world settings. The robot begins the interaction and image description generation process upon being verbally asked by the user or a tap on the robot’s head.

Although the NAO robot has an incredibly powerful CPU processor, it does not have the capacity to run complex models like the proposed system in real time. To this end, the robot acts essentially as a front end that interacts with the system deployed on a more powerful GPU server. The processing procedures are as follows. The robot captures an image from its cameras, sends it via a LAN to the remote server to be processed, then receives a response in the form of raw text and finally verbally outputs the generated description of the captured image using its text-to-speech API. Preliminary experimentation shows that the robot is

capable of observing, recognizing and describing diverse objects (such as cups, fruits, furniture etc) and people, as well as their attributes within multiple environments, such as ‘stairway’, ‘library’, ‘kitchen’, or ‘office’. The server used throughout the experiments is based on the Nvidia Deep Learning DevBox [29] equipped with 4 GTX TITANS. We will also conduct more experiments to explore the efficiency of the proposed system integrated with the NAO robot for diverse real-life settings as one of the future directions.

As an initial indication of the system efficiency, we deploy the robot platform integrated with the proposed deep network to a real-life application scenario. Table 2 shows images fed through the robot’s network and the aspects in which the robot reports based on image description generation output. The empirical results indicate the efficiency of the proposed system in dealing with real-life images via the robot platform. In future work, we aim to incorporate floor detection methods with the existing object and scene recognition to allow the system to detect hazards and audibly present this information via the robot platform to benefit e.g. healthcare application scenarios. We also aim to equip the proposed system with the capabilities of dealing with low resolution images to further enhance performance.



## 4. Evaluation

In order to evaluate the efficiency of the proposed system, we implement two popular baseline methods, i.e. Google NIC [1] and NeuralTalk [2], for comparison. The IAPR TC-12 [23] dataset has been used for evaluation. The IAPR TC-12 dataset consists of 20,000 images, with each image paired with one descriptive sentence or a short paragraph. In the test stage, we ran our system on a random selection of ~10,000 images.

We have trained our RNN-based language models purely on two small caption subsets of IAPR TC-12 and MSCOCO, respectively, without using the associated images. That is, we only use a small number of captions from each dataset for training and the training process has not used any images from either of these datasets. The baseline methods, i.e. NeuralTalk and Google NIC, however, have been trained on these datasets (using both captions and images), therefore leading to higher evaluation scores. We still provide the evaluation results using these datasets in order to indicate the efficiency of the proposed model.

To quantify the performance of the system, the MSCOCO

**Table 2.** Example images and outputs produced by the system deployed on the NAO robot

	
<b>Description output</b>	“two women and six men are sitting behind a wooden table in a room with a light yellow wall”
	“a man and three women are walking on a slope with a white ladder and bushes behind them, snow covered mountain range in the background”

**Table 3.** Our results and comparison with related work on the IAPR TC-12 dataset

IAPR TC-12 [23]	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SPICE
<b>Our Work</b>	<b>0.201</b>	<b>0.105</b>	<b>0.053</b>	<b>0.024</b>	<b>0.073</b>	<b>0.216</b>	0.038
<b>NeuralTalk</b>	0.129	0.069	0.038	0.022	0.065	<b>0.216</b>	<b>0.061</b>
<b>Google NIC</b>	0.094	0.046	0.034	0.013	0.059	0.205	0.06

evaluation script, is utilized. The MSCOCO script contains four BLEU metrics [30] (i.e. BLEU-1, BLEU-2, BLEU-3 and BLEU-4) based on the  $n$ -gram method of determining string/sentence similarity. It is also equipped with other evaluation metrics such as, METEOR [31], ROUGE-L [32], and SPICE [33]. The detailed results using all of the above metrics for the evaluation of the IAPR TC-12 dataset are shown in Table 3. We also illustrate example images from the test dataset, along with their generated paired descriptions in Fig. 4.

The proposed architecture is designed and motivated to expand upon the short captions produced by existing research. As indicated in Table 3, our largest improvement over existing work is in the BLEU-1 metric. This could be attributed to the addition of attribute prediction, or the prediction of more attributes, in comparison with those of the existing methods. On top of this, generating individual words that are more likely to be present in the reference sentence would increase the lower  $n$ -gram BLEU score (e.g. BLEU-1), which takes the frequency of the words and the length of the description into account for score generation.

On the other hand, generating words or attributes, which are correct, but may not be present within the reference description, reduces the score within the higher  $n$ -gram metrics, such as BLEU-3 or BLEU-4. This effect is most noticeable within the ROUGE-L score. Our system has the capability to generate multiple attributes for a given object. If the reference description only contains one or two attributes, and our system generates more than that, our score in the higher  $n$ -grams would be penalized. A simplified example is given below, which illustrates example theoretical descriptions generated by the proposed model and a typical existing framework such as Google NIC.

The proposed model: “a young man wearing a red striped shirt”

An existing method: “a man wearing a shirt”

Reference: “a man wearing a red shirt”

The BLEU-1 score is calculated for each word, so, for example, each generated word would score a precision of  $x/($ the length of the corresponding generated sentence by a specific method), depending upon its frequency. Only the word ‘red’ in the above example of the proposed model would receive the precision score of this word, whereas the existing method would receive 0 owing to the fact that the attribute ‘red’ is not generated by the existing method. Therefore, our model scores well for the BLEU-1 metric, however the score suffers with the higher  $n$ -grams, since the correct ‘red striped’ does not appear in the reference. This is attributed to the nature in which our system is trained and built. Having trained separately and solely in sections of large scale attribute datasets, this enables our model to put more focus into attribute annotating than existing methods, even if some attributes are not present in the reference corpus.

#### 4.1. Experimental Results

The empirical results for the evaluation of the IAPR TC-12 dataset indicate that the proposed system outperforms the two baseline methods of similar structures. The BLEU-1 score obtained by all the methods is lower than the human performance of around 0.6, however this is to be expected due to the cross-dataset evaluation. The BLEU metric provides an insight into the similarity comparison of the words’ and short phrases’ levels. The higher levels of BLEU metrics indicate the comparison of longer strings in the source and target sentences.

The comparatively higher scores throughout all of the BLEU metrics show how our work outperforms its competition. Specifically, for the BLEU evaluation, our work outperforms NeuralTalk by an average of 0.035 and NIC by an average of

0.053. The proposed system also outperforms the two baseline methods for the METEOR measures. In comparison to BLEU, the METEOR metric has gained increasing popularity owing to the closer correlation between the sentence level information, to human performance.

Moreover, the ROUGE-L metric is also used for evaluation. This metric is looking for the longest matching subsets between the automatically generated captioning and the human annotation. Although the captions generated by the different methods show great distinctions, the three systems achieve identical ROUGE-L scores.

The SPICE metric is also used for evaluation. It determines the semantic similarity between the pair of a generated description and its ground truth annotation. The proposed system achieves the lower score for the SPICE metric, in comparison to those of other metrics. This could be caused by the lengthy descriptions generated by the proposed system which may challenge and affect the semantic similarity score calculation in SPICE. SPICE utilizes scene graphs. The longer reference and generated descriptions could make a greater difference between these graphs, making a high similarity harder to achieve.

Overall, our proposed deep network outperforms systems of a similar structure, when all methods have been tested on different images to their training sets. This shows that our system has sufficient diversity and possesses the ability to generate descriptive captions for real-life and staged images. As can be seen in Fig. 4, our results are also considerably longer and more descriptive, and in many cases correct, in comparison to those generated by related methods.

## 5. Conclusion and Future Work

In this research, we have proposed a novel deep network architecture for region annotations and full image description generation. The proposed model consists of a set of deep networks, including the regional proposal generator, CNNs and RNN-based encoder-decoder, to achieve a high level of quality for image description generation. By employing a regional approach, the proposed system is able to collect, annotate and describe a large number of details overlooked by other typical methods. It also requires dramatically fewer training images. The proposed framework has also shown its significance in dealing with out-of-domain datasets, which challenge other state-of-the-art methods significantly, as shown in the evaluation of the IAPR TC-12 dataset. The overall architecture of our model is complex, combining multiple techniques and procedures to deliver effective image description generation. In future work, exploration in reducing the number of layers, model stages, and the feature complexity will be conducted to improve the system efficiency and runtime, and potentially the results.

In future work, we aim to explore another advanced deep network, i.e. Generative Adversarial Networks (GANs) [34], because of its superior capability for image generation. We aim to explore its adaptation for image description generation owing to its unique style of training. Specifically, GANs are composed of two models i.e. the generative and discriminative models, which are trained simultaneously. The latter estimates that some data belongs to the training set, or some generated by the generative model. The generative model is trained to maximize the probability of the discriminative model making a mistake. Such training mechanisms may benefit image description generation tasks as well. We also intend to incorporate an attention or saliency mechanism into the region proposal generation stage to improve upon the quality of the generated regions of interest.

Finally, we also aim to produce a large-scale image description dataset that is more descriptive and discriminative than existing publicly available datasets. This would allow for not only a more accurate representation and evaluation of our model, but also further research into the descriptive caption generation rather than the typically available short captions.

## References

- [1]. O. Vinyals, A. Toshev, S. Bengio and D. Erhan, 2015. Show and tell: A neural image caption generator. In Proc. IEEE conf. Computer Vision and Pattern Recognition. Boston, Massachusetts. 3156-3164
- [2]. A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proc. IEEE conf. Computer Vision and Pattern Recognition. Boston, Massachusetts. 3128-3137.
- [3]. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- [4]. T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, (2010, September). Recurrent neural network based language model. In *Interspeech* (Vol. 2, p. 3).
- [5]. A. Graves and J. Schmidhuber, 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures". *IEEE Trans. Neural Netw.* 18. 602-610
- [6]. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. Unpublished. arXiv preprint arXiv:1412.3555.
- [7]. S. Ren, K. He, R. Girshick and J. Sun., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99)
- [dataset] [8]. P. Young, A. Lai, M. Hodosh, J. Hockenmaier, 2014, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", *Transactions of the Association for Computational Linguistics*. 67-78.
- [dataset] [9]. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, 2014. Microsoft coco: Common objects in context". In *Computer Vision–ECCV 2014* (pp. 740-755).
- [dataset] [10]. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidi, L. Li, D.A. Shamma, M. Bernstein, L. Fei-Fei, 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Unpublished. Available at: <https://arxiv.org/abs/1602.07332>
- [11]. K. Xu, J.L. Ba, R. Kiros, K. Cho et al, 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In Proc. Conf. International Conference on Machine Learning. 2048 - 2057
- [12]. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg and T.L. Berg, (2011). Baby talk: Understanding and generating image descriptions. In *Proceedings of CVPR*.
- [13]. J. Johnson, A. Karpathy, and L. Fei-Fei, 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4565-4574
- [14]. Y.H. Tan, and C.S. Chan, 2016. phi-LSTM: A Phrase-based Hierarchical LSTM Model for Image Captioning. arXiv preprint arXiv:1608.05813
- [15]. E. Matsuo, I. Kobayashi, S. Nishimoto, S. Nishida, and H. Asoh, (2016). Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity. *ACL 2016*, 22
- [16]. H. Fang, S. Gupta, F. Indola, R. Srivastava et al. 2015. From Captions to Visual Concepts and Back". In *Proc. IEEE conf. Computer Vision and Pattern Recognition*. Boston, Massachusetts. 1473-1482
- [17]. K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler and C. Sienkiewicz, 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 49-56)
- [18]. Y. Sugano and A. Bulling, 2016. Seeing with humans: Gaze-assisted neural image captioning. Unpublished. arXiv preprint arXiv:1608.05203
- [dataset] [19]. D. Elliott, S. Frank, K. Sima'an, and L. Specia, 2016. Multi30K: Multilingual English-German Image Descriptions. arXiv preprint arXiv:1605.00459. Unpublished
- [20]. A. Krizhevsky, I. Sutskever and G.E. Hinton, 2012. Imagenet classification with deep convolutional neural networks". In *Advances in neural information processing system*. 1097-1105
- [dataset] [21]. O. Russakovsky and L. Fei-Fei, 2010. Attribute learning in large-scale datasets". In *Trends and Topics in Computer Vision*. 1-14
- [dataset] [22]. N. Kumar, A.C. Berg, P.N. Belhumeur and S.K. Nayar, 2009. Attribute and simile classifiers for face verification". In *Proc. IEEE Int. Conf. Computer Vision*. 365-372.
- [dataset] [23]. M. Grubinger, P. Clough, H. Müller and T. Deselaers, 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems". In *Int. Workshop OntoImage* (5)10.
- [dataset] [24]. J. Xiao, J. Hays, K A. Ehinger, A. Oliva and A. Torralba, 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition*. 3485-3492.
- [25]. D. Bahdanau, K. Cho, and Y. Bengio, 2014. Neural machine translation by jointly learning to align and translate". arXiv preprint arXiv:1409.0473. Unpublished.
- [26]. P. Koehn, 2009. *Statistical machine translation*. Cambridge University Press
- [27]. I. Sutskever, O. Vinyals and Q.V. Le, 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112)
- [28]. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
- [29]. NVIDIA DIGITS DevBox. Available: <https://developer.nvidia.com/devbox>. Last accessed 25th Jan 2017.
- [30]. K. Papineni, S. Roukos, T. Ward and W.J. Zhu, 2002. BLEU: a method for automatic evaluation of machine translation". In *Proc. of the 40th Annual meeting on association for computational linguistics*. ACL. 311-318.
- [31]. S. Banerjee and A. Lavie, 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In *Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. (29) 65-72.
- [32]. C.Y. Lin, 2004. Rouge: A package for automatic evaluation of summaries". In *Text summarization branches out: Proc. of the ACL-04 workshop* (8)
- [33]. P. Anderson, B. Fernando, M. Johnson and S. Gould, 2016, October. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision* (pp. 382-398)
- [34]. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, Warde-D. Farley, S. Ozair, A. Courville and Y. Bengio, 2014. Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).