Generating Descriptive Paragraphs For Images Using Coherence Vectors

Rakshit C S, Dept of ECE, PES University, **Sirajahamed N D**, Dept of CSE, PES University
Prof Rama Devi P, Dept of CSE, PES University

Abstract

Image Captioning is the process of generating textual description of an image. It uses both **Natural Language Processing** and **Computer Vision** to generate the captions. Whereas **Paragraph generation from images** is a similar process with larger sentences which is an important task for video summarization, editing, and support of the disabled, which has attracted extensive research attention recently. As the traditional image captioning process doesn't serve this purpose we present an approach to generate descriptive paragraphs for the image using **coherence vectors** and **global topic vectors**.

Introduction

We daily capture the fun moments of our life and describe those events to our friends and family showing them the pictures, well computer vision and Natural Language Processing makes life easy by generating automated captions and paragraphs which give a description of the picture. With the advent of large datasets pairing images with natural language descriptions such as MS COCO dataset, Visual Genome Dataset, etc, it has recently become possible to generate novel sentences describing images.

Image captioning uses deep learning techniques such as CNNs and RNNs to describe images. The task of image captioning can be divided into two modules logically – one is an **image based model** – which extracts the features and nuances out of our image, and the other is a **language based model** – which translates the features and objects given by our image based model to a natural sentence. For our image based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language based model (viz decoder) – we rely on a Recurrent Neural Network.

In this paper we address the limitation of image captioning, i.e. the context of the scene by generating descriptive paragraphs for the given image using a methodology called as Coherence Vectors and Global topic vectors. The coherence vector has the context of the recently generated sentence and when this vector is coupled with the topic vector of the current generating sentence gives a meaningful paragraph and the context is preserved. We test the model on a mixture of MS COCO dataset and Visual genome dataset. This dataset contains a total of 19575 images with descriptive paragraphs.

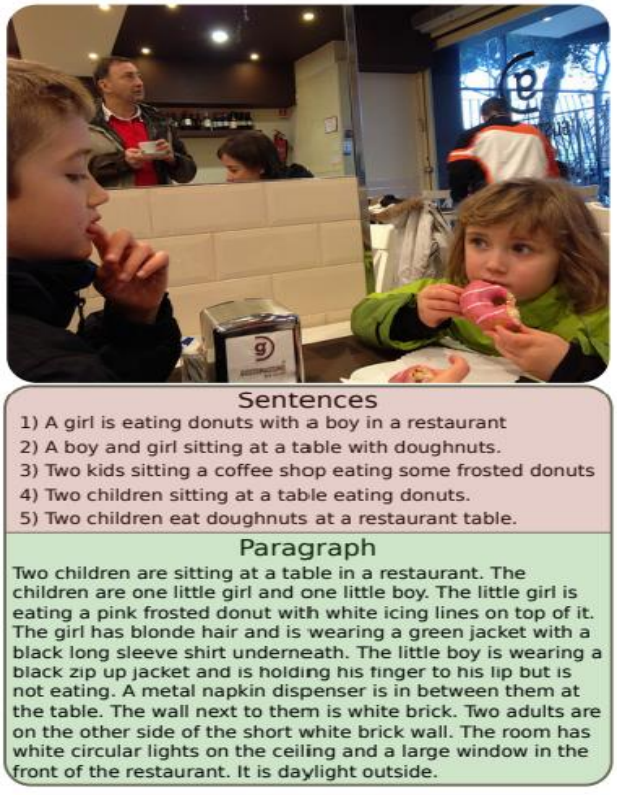

Fig 1: Difference between captions and paragraphs for the same image.

Literature Survey

There have been multiple approaches for image captioning such as underwater image captioning as in [1] . In this paper they use an image based model and LSTMs in language based model. But all image captioning models fail to give detailed description for an image. There are a few paragraph generation models such as in [2] . The method described in this paper uses dense captioning technique to generate sentences (captions) and hierarchical RNN to generate paragraphs from those sentences.

Proposed Method for Paragraph Generation

Any image captioning model is divided into 2 parts as in fig 2. An **Image based model** and a **Language based model**. The image based model is used to generate feature vectors of the images. This process is also known as feature extraction. At the end of this process the **Pooled Visual Feature Vector** is obtained which can be further fed into the language model to generate sentences. The **Language based model** consists of Deep network which generated sentences based on the feature vectors.

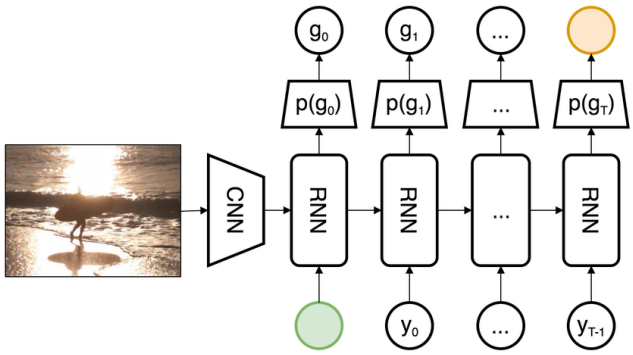

Fig 2: Architecture of a common image captioning model.

Image based model

This is a **Deep network** model which uses **CNN** based architecture to extract the features of the given image. In our prescribed method we have used **Google's InceptionV3** model which has a series of Convolution layers, pooling layers and activation layers. The image model takes input **image x of size (299, 299, 3)** and returns the **feature vector of size 2048**. This network gives out different **regions of interest** based on different filters used by the convolution layer. These regions are projected onto the **convolutional feature map** consisting of many layers and then this is connected to a fully connected layer to give a vector of size 2048.

Language based model

To generate a paragraph $\mathbf{Y = (Y_1 \ldots Y_S)}$ Consisting of S sentences $\mathbf{Y_i, \forall\, i \in \{1 \ldots S\}}$, each with Ni words $\mathbf{Y_{i,j}, \forall\, j \in \{1 \ldots Ni\}}$, for an **image x**, we make use of a deep net composed out of two modules which are coupled hierarchically: the **Topic Generation Net** and the **Sentence Generation Net**.

The **Topic Generation Net** extracts a set of S **topic vectors**, $\mathbf{Ti \in R^H\ \forall\, i \in \{1 \ldots S\}}$, given an appropriate visual representation of the input image x. The topic generation net is a parametric function which produces a topic vector Ti and a probability measure ui indicating if more topics are to be generated. We implement this function using a recurrent net, referred to as the **SentenceRNN**. We then combine all the topic vectors Ti to construct a **Global Topic Vector** $\mathbf{G \in R^H}$, which captures the underlying image summary. This global topic vector is constructed via a **weighted combination** of the topic vectors Ti. Formally we use $\mathbf{(G, \{(Ti, ui)\}\ S\ i=1) = \Gamma wT\ (x)}$ to denote the input and output of the net $\mathbf{\Gamma wT\ (\cdot)}$, where the vector **wT** subsumes the parameters of the function. This is illustrated in fig 3.

The second module of this approach is the **sentence generation net.** Based on the output of the Topic generation net this module generates sentences Yi at once to produce a paragraph Y. Formally, the sentence generation module is also modeled as a parametric function which synthesizes a sentence Yi , one word $Y_{i,j}$ at a time. More specifically, a recurrent net $\Gamma ws\ (\cdot, \cdot)$ is used to obtain the **predicted word probabilities** $\mathbf{\{pi,j\}\ Ni\ j=1 = \Gamma ws\ (Ti, G)}$, where ws subsumes all the parameters of the net, and $\mathbf{pi,j \in [0, 1]V\ \forall j \in \{1, \ldots, Ni\}}$ is a probability distribution over the set of V words in our vocabulary. We realize the function, $\Gamma ws\ (\cdot, \cdot)$

using a recurrent net, subsequently referred to as the **WordRNN.**This is illustrated in fig 4.

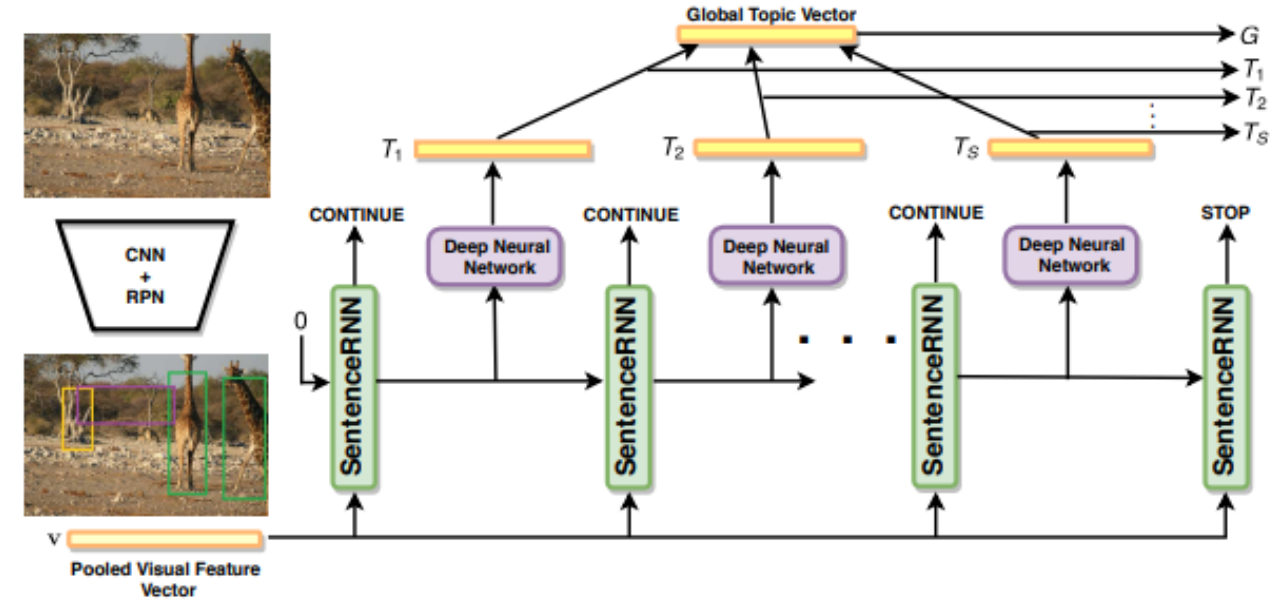

Fig 3: Illustration of Topic Generation Net from Pooled visual feature vector using SentenceRNN

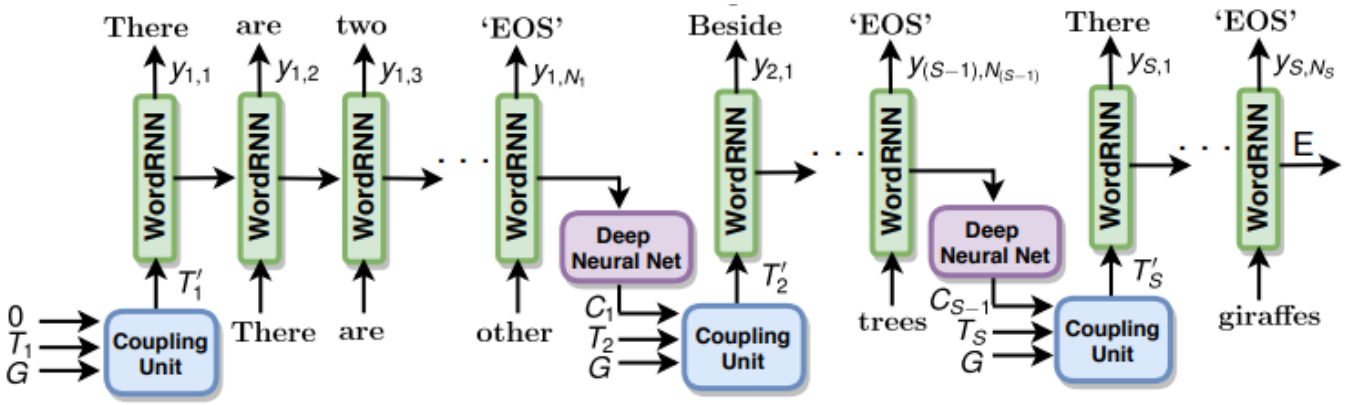

Fig 4: Illustration of paragraphs from topic vectors Ti and coherence vectors C_{i-1} using WordRNN.

Coherent paragraph generation

Topic Generation Net: The visual pooled vectors ***v*** is used the input to the sentenceRNN. The sentenceRNN has a **single layer GRU (Gated Recurrent Unit)** which produces **'CONTINUE'** and **'STOP'** labels with the probability **ui** and its hidden state is used to produce the topic vectors $\mathbf{Ti} \in \mathbf{R^H}$. We construct the Global topic vector as shown in fig 3. The global topic vector is as follows:

$G = \sum_{i=1}^{n} \alpha i Ti$ where $\alpha i = \frac{||Ti||2}{\sum_i ||Ti||2}$

Sentence Generation Net: For each of the topic vector Ti one sentence is generated. Synthesis of each sentence Yi begins by combining the coherence vector of the previous sentence, C_{i-1}, with the current topic vector, Ti, via a coupling unit.

The coupling unit produces a modified topic vector Ti' which is propagated to the WordRNN which has GRU's to generate sentences. The sentence is terminated either when the max limit is reached or the 'EOS' tag is predicted by the wordRNN.

Coherence Vectors: An important element of human-like paragraphs is coherence between the themes of successive sentences, which ensures a smooth flow of the line of thought in a paragraph. In order to compute the coherence vector for the (i − 1)th sentence, we extract the hidden layer representation (∈ R H) from the WordRNN, after having synthesized the last word of the (i − 1)th sentence. This encoding carries information about the (i−1)th sentence, and if favorably coupled with the topic vector Ti of the i th sentence, encourages the theme of the i th sentence to be coherent with the previous one.

Coupling Unit: Having obtained the coherence vector Ci−1 from the (i − 1)th sentence, a Coupling Unit combines it with the topic vector of the next sentence, Ti , and the global topic representation G. This process is illustrated in fig 5. More specifically, we first combine C_{i-1} and Ti into a vector $T_i^C \in R\ H$ which is given by the solution to the following optimization problem:

$$T_i^{\,c} = argmin_{T_i \hat{}\, c}\ (\alpha \left|\left|T_i^{\,c} - T_i \hat{}^{\,C}\right|\right|_2^{\,2} + \beta \left|\left|C_{i-1} - T_i \hat{}^{\,C}\right|\right|_2^{\,2})\ with\ \alpha, \beta \geq 0$$

The solution to this equation is:

$$T_i^{\,c} = \frac{\alpha(T_i) + \beta(C_{i-1})}{\alpha + \beta}$$

$$for\ both\ \alpha, \beta\ not\ equal\ to\ 0$$

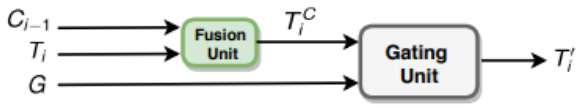

Fig 5: Illustration of coupling unit.

Loss function and Training

There are 2 loss functions, one for the Topic generation net and one for Sentence generation net. We have use **mean squared error** as the loss function for **Topic Generation net** and **Cross Entropy loss** for **Sentence Generation net.**

The train process is simple. In the forward process for each image and respective paragraph, the global topic vector is found out and the loss is calculated. Further using these topic vectors, the sentences are formed and the loss is calculated correspondingly. These 2 loss values are added and the gradients are calculated in the backward network. **Adam optimizer** is used as the optimizer for the network.

We have trained the network on the image dataset as mentioned earlier.

Experiment Evaluations

We report the performance of our model based on the BLEU scores.

Outputs achieved are as follows:

GENERATED OUTPUT FOR THE IMAGE

is giraffe standing in the middle of the street and.is large forest of trees in the background and the.giraffe has its head down the brown tree and the.giraffe has brown and tan horns on its head and.is small tuft of water in the background that is.is tree behind the car that is green with white.giraffe has its head down the brown and tan brown.is on the bottom half of the bus and the

Bleu Scores :

Bleu 1	0.8163
Bleu 2	0.1101
Bleu 3	0.0169
Bleu 4	0.0045

References

[1] **Generating Captions for Underwater Images Using Deep Learning Models**: Hardik Gourisaria, Shylaja S S, Rama Devi P, Akhilarka Jayanthi, Tanay Gangey Department of Computer Science and Engineering, PES University, Bangalore, India.

[2] **A Hierarchical Approach for Generating Descriptive Image Paragraphs**: Jonathan Krause Justin Johnson Ranjay Krishna Li Fei-Fei Stanford University.

[3] **Diverse and Coherent Paragraph Generation from Images**: Moitreya Chatterjee and Alexander G. Schwing University of Illinois at Urbana-Champaign, Urbana IL 61801, USA