Major Project Proposal Report

On

Image Captioning with Region-Based Attention and Scene-Specific Contexts

Submitted by

15IT111 VIRANCHI BADHEKA
15IT132 NIHAR RAICHADA
15IT231 RIA KULSHRESTHA
15IT243 SHIVANI SHRIVASTAVA

Under the Guidance of

Mr. Dinesh Naik

Dept. of Information Technology,
NITK, Surathkal

Date of Submission: 8th October, 2018

Department of Information Technology
National Institute of Technology Karnataka,
Surathkal.
2018-2019

Abstract

A region based attention model that automatically learns to describe the content of images using Recurrent Neural Networks and Scene-Specific contexts. The image captioning system will exploit the parallel transitioning dynamics between the visual focus and the sentences. Scene-specific contexts will capture higher-level semantic information encoded in an image. The contexts will adapt to language models for word generation to specific scene types The performance of the model is bench marked on the MS COCO dataset. Deep network attention can be viewed as a form of alignment from language space to image space. These attention maps carry important information in understanding deep networks.

Keywords: Image captioning, visual attention, scene-specific context, LSTM.

Contents

Abstract i

1 Introduction 1

1.1 Motivation 1

2 Requirement Analysis 2

2.1 Literature Survey 2

2.2 Outcomes of Literature Review 5

2.3 Problem Statement 6

2.4 Objectives and Methodology 6

References 10

1 Introduction

1.1 Motivation

Automatically generating captions of an image is a task very close to the heart of scene understanding and is one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language.

One of the most curious facets of the human visual system is the presence of attention. Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. One advantage of including attention is the ability to visualize what the model sees. We implemented a model that can attend to salient part of an image while generating its caption.

2 Requirement Analysis

2.1 Literature Survey

[7] proposed the idea of a neural and probabilistic framework to generate descriptions from images. They made use of a recurrent neural network(RNN) which encodes the variable length input into a fixed dimensional vector, and uses this representation to decode it to the desired output sentence.
But they replaced the encoder RNN by a deep convolution neural network (CNN). As CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector.

[1] reveals a significant limitation of simple RNN models which strictly integrate state information over time known as the vanishing gradient or exploding gradient problem which makes it really hard to learn and tune the parameters of the earlier layers in the network and worsens as the number of layers in the architecture increases.
LSTMs solve this by incorporating memory units that explicitly allow the network to learn when to forget previous hidden states and when to update hidden states given new information.
They proposed adding additional depth to LSTMs by stacking them on top of each other and using the hidden state of the LSTM in current layer as the input to the LSTM in next layer.
They show stacked LSTM models provide improved recognition on conventional video activity challenges and enable a novel end-to-end optimizable mapping from image pixels to sentences.
They use an multimodal encoder-decoder model based on LSTM networks. They used a visual conv-net to encode a deep state vector, and an LSTM to decode the vector into a natural language string.

[6] proposed a multi-modal Recurrent Neural Networks (m-RNN) model to address both the task of generating novel sentences descriptions for images, and the task of image and sentence retrieval. The whole m-RNN architecture contains a language model part which learns the dense feature embedding for each word in the dictionary and stores the semantic temporal context in recurrent layers, an image part which contains a deep Convulutional Neural Network (CNN) that extracts image features and a multimodal part which connects the language model and the deep CNN together by a one-layer representation.

[2] approaches to caption generation by incorporating the attention mechanism. Rather than compressing an entire image into a static representation, attention mechanism allows for salient features to dynamically come to the forefront as needed. This is useful when there is a lot of clutter in an image.
The paper proposes the Encoder-Decoder framework.Encoder is responsible for extraction of visual features from corresponding 2D portions of the images.The extraction is done from lower convolutional layers that allows selective focus on certain parts of an image by selecting a subset of all the feature vectors.
Decoder uses LSTM that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. For each location 'i' extracted by encoder, the mechanism generates a positive weight that represents the probability that location 'i' is the right place to focus for producing the next word, hence finding the region of interests.

[3] proposes the algorithm that combines the bottom-up and top-down approaches of image captioning through a semantic attention model.
The semantic attention model attends a semantically Region of Interest(ROI) in an image,weighs the relative strength of attention paid on multiple ROIs and is able to switch attention among ROIs dynamically.
The topdown paradigm starts from a visual feature representation of an image and converts it into words, The botton-up paradigm shows the changes of the attention weights for several candidate ROIs with respect to the recurrent neural network iterations.
The model is built on top of a Recurrent Neural Network (RNN), whose initial state captures global information from the top-down feature. As the RNN state transits, it gets feedback and interaction from the bottomup attributes via attention mechanism.
The feedback allows more accurate and robust inferences of semantic gap between generated caption and actual image content.

[4] proposes "Selective Search" by using segmentation to generate a limited set of locations(Regions of Interest), i.e. best locations for object recognition. they put most emphasis on recall and good object approximations over exact object boundaries. They used a full segmentation hierarchy and accounted for as many different scene conditions as possible, such as shadows, shading, and highlights, by using a variety of invariant colour spaces.
They started with oversegmentation the image, i.e. a set of small regions which do not spread over multiple objects. Starting from the initial regions, they used a greedy algorithm which iteratively groups

the two most similar regions together and calculates the similarities between this new region and its neighbours. This process is continued until the entire image becomes a single region.
Either all segments throughout the hierarchy (including initial segments) or the tight bounding boxes around these segments are considered as potential object locations.
The similarity between any two regions is defined as a sum of the fraction of the image that the segment a and b jointly occupy and defined as the histogram intersection between SIFT-like texture measurements. Both the factors are equally weighted.

[5] proposes a generative model: Latent Dirichlet Allocation (LDA) that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posts that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

2.2 Outcomes of Literature Review

RNNs were initially proposed for caption generation but they lacked scalability due to their vanishing gradient or exploding gradient problem which worsened as the depth of the architecture increased.

This problem was solved by LSTMs which incorporated memory units that explicitly allowed the network to learn when to forget previous hidden states and when to update hidden states given new information. Also, stacking multiple LSTMs in a model and adding a feedback mechanism between the layers improves recognition and enables end -to-end mapping from input image to sentences.

The Encoder-Decoder framework is implemented by successive CNN and RNN sub models.CNN extracts the visual features from the image and feeds it to RNN for the sequential generation of sentence describing the image.

Attention Mechanism further enhances the accuracy and robustness of the models by focusing on the salient regions of the image. By using selective search to generate ROIs we get higher recall, coarse locations of objects and less computations.

LDA allows capture of higher-level semantic information encoded in an image.Being a generative probabilistic model of a corpus, it allows representation of documents as random mixtures over latent topics, where each topic is characterized by a distribution over words.

2.3 Problem Statement

To model an image captioning system that understands the image, recognizes the objects in it, reasons about the relationship among those objects, and focuses on the more salient parts in the image by exploiting the close correspondence between visual concepts(detected as object-like regions) and their textual realization as words in sentences.

2.4 Objectives and Methodology

- To generate localized candidate regions at multiple scales, which contain visually salient objects, to represent images.

Methodology: We use the technique of selective search to construct a hierarchical segmentation of the image. The technique first uses color and texture features to over-segment the image, and merges neighboring regions to form a hierarchy of segmentations until the whole image merge into a single region.
Among the large number of regions at different levels (i.e.,scales), we select semantically meaningful, primitive and non-compositional but contextually rich ones by training a binary classifier which indicates whether a region is good or bad.

- To represent the above generated images as a collection of visual feature vectors computed on localized regions at multiple scales.

  Methodology: We use the ground-truth bounding boxes as positive examples, and randomly select some patches, which have at most 30 percent intersection with ground-truth, as negative examples. A logistic regression model is trained to classify the patches and applied to new images (including those from other datasets). The classifiers outputs are considered as a measure of objectness how likely it is for an image region to contain object of any class. Since the objects have diverse scales, the top regions scored by the classifier are diverse in sizes.We resize each patch into 224x224 to feed into the ResNet-152 network to obtain 2,048-dimensional CNN features.
  For each image, we select the top 29 regions according to the outputs of the classifier, and additionally include the whole image to have in total R=30 regions. Diverse sizes are preferred as not all objects in a image have the same size.

- To build an LSTM-based neural network that models the attention dynamics of focusing on those regions as well as generating the words sequentially.

  Methodology: We hypothesize there is a process $\{\mathrm{h_t}\}$ of latent meaning, governing the transitions from one concept to another. In our models, we have three sets of variables: the hidden state $\{\mathrm{h_t}\}$ for characterizing the transition of latent meanings, the output variables$\{\mathrm{w_t}\}$ for the words being generated, and the input variables $\{\mathrm{v_t}\}$ describing the visual context for the image, for example, for the visual element(s) being focused.

  For simplicity, the subscript t indices the time, expanding from 0 (START) to T + 1 (END) where T is the length of the sentence. Our model is a sequential model, predicting the value of the new state and output variables at time t + 1, based on their values in the past as well as the values of input variables up to time t+1.

  At any time t, our system is presented with the image which is already analyzed and represented with R localized patches at multiple scales. At time t, we predict which visual element is being focused and obtain the right feature vector as visual context by using an one-hidden-layer neural network with R softmax output variables.

- To extract scene-related global contexts, followed by injecting scene contexts into LSTMs.

  Methodology: The ideas used will be : 1) Unsupervised clustering of captions into scene categories and 2) Supervised learning of a regressor to predict the scene vectors from the visual appearance.

  Unsupervised Clustering : For each image, a multi-dimensional topic vector is obtained that assigns its caption into the memberships of multiple categories , called scene vectors.Scene vectors are inferred from the corpus that is generated using the captions of training dataset.

Supervised Learning : The clustering is followed by the training of a regressor for prediction of scene vector when presented with an image. The training samples for this regressor are the images from the training dataset with the target outputs being the inferred scene vectors. A multilayer perceptron will be used as a regressor.In case when the captions are not available, global feature vectors of the images will be used.

The scene vectors will be then injected into LSTM and the sentence generation process will adapt to be scene-specific by biasing the input gates and the input modulator in the LSTM.The task of the modulator is to ensure constant bias during sentence generation.

References

[1] J. Donahue, et al., Long-term recurrent convolutional networks for visual recognition and description, in Proc. IEEE Comput. Vis. Pattern Recognit., 2014, pp. 26252634.

[2] Chen, X., and Zitnick, C. L. 2014. Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654.

[3] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders,Selective search for object recognition, Int. J.Comput. Vis.,vol. 104, no. 2, pp. 154171, 2013.

[5] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, Explain images with multimodal recurrent neural networks, in Proc. Int. Conf.Learn. Representations, 2015.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., 2015, pp. 31563164.

[7] . Koch Biophysics of Computation: Information Processing in Single Neurons. New York:: 1998.