Image Captioning with Convolutional Neural Networks

Makwana sudhir

18IT057

Chandu bhai s.patel institute of technology, charusat

Changa,anand

18it057@charusat.edu.in

Prof . Amit parmar

Chandu bhai s.patel institute of technology

charusat

Changa,anand

amitparmar.It@charusat.ac.in

Abstract

In this thesis, we elaborate on image captioning concerning especially dense image captioning. We present technical fundamentals of a model striving to solve such a task. Concretely, a detailed structure of DenseCap and Neural Image Caption is discussed. Experimentally, we examine results of DenseCap and analyse the model's weaknesses. We show that 92% of the generated captions are identical to a caption in the training set while the quality of those and the novel

ones remains the same. We propose a criterion that significantly reduces a set of captions addressing an image whilst SPICE score of the set is maintained.

Keywords: image captioning, dense captioning, convolutional neural networks, long short-term memory

Image Captioning with Convolutional Neural Networks

INTRODUCTION

Every day we see a lot of photographs in the surroundings , on social media and in the newspapers. Humans are able to recognize photographs themselves only. We humans can pick out the photographs without their designated captions but on the other hand machines need images to get trained first then it'd generate the photograph caption automatically.
Image captioning may benefit for loads of purposes, for example supporting the visionless person using text-to-speech through real time feedback about encompassing the situation over a camera feed, improving social medical leisure with the aid of reorganizing the captions for photographs in social feed alongwith messages to speech.

Facilitating kids in recognizing substances further to gaining knowledge of the language. Captions for every photograph on the world wide web can produce quicker & detailed authentic photographs exploring and indexing. Image captioning has diverse packages in numerous fields inclusive of biomedicine, commerce, internet looking and navy and many others. Social media like Instagram , Facebook etc can generate captions routinely from images
The principal goal of this research paper is to get a little bit of expertise in deep learning strategies. We use two strategies specially

Image Captioning with Convolutional Neural Networks

(2) Neural Networks

In order to tackle the image captioning task, recent work shows it is in one's interest to utilize neural networks . This frequently used term dates back to 1950s when notions such as the Perceptron Learning Algorithm were introduced. Modern neural networks draw on notions discovered in the era of a Perceptron. In this section, we first define a neuron as a fundamental part of modern neural networks. Then we elaborate on Convolutional Networks and Recurrent Networks

2.1 Perceptron

For the purposes of this work, a perceptron is defined generally as it became a fundamental part of modern neural networks and the notation is utilized further on. Thus, a perceptron is compounded of one neuron. The neuron's output, known as the activation a, is mapped by $\Phi : R N \rightarrow R$ as follows:

$$a = \Phi(x) = \sigma(wT x + b) \qquad (1)$$

where $x \in R N$ is a feature vector, $w \in R N$ and $b \in R$ are weights and $\sigma(\cdot)$ is a non-linear function. In case of the Perceptron, $\sigma(\cdot)$ stands for

2.2 multi layer neural network

A multi-layer neural network **contains more than one layer of artificial neurons or nodes**. They differ widely in design. It is important to note that while single-layer neural networks were useful early in the evolution of AI, the vast majority of networks used today have a multi-layer model.

Deep Learning

1. Deep Learning is a prime technology behind the technology such as virtual assistants, facial recognition, driverless cars, etc.
2. The working of deep learning involves training the data and learning from the experiences.
3. The learning procedure is called ‘Deep’, as with every passing minute the neural networks rapidly discover the new levels of data. Each time data is trained, it focuses on enhancing the performance.
4. With the increasing depth of the data, this training performance and deep learning capabilities have been improved drastically, and this is because it is broadly adopted by data experts.

Image Captioning with Convolutional Neural Networks

2.4 Convolutional Neural Networks

In image analysis, many of recent advances in deep learning are built on the work of LeCun et al. who introduced a Convolutional Neural Network (CNN) which had a large impact on the field. A CNN is a type of a neural network that is designed to process an image and represent it with a vector code. The architecture of CNNa draws on fully-connected neural networks. Similarly, a convolutional neural network is a compounded structure of several layers processing signals and propagating them forward.

Convolution in CNNs

A neuron's receptive field is processed similarly to fully connected layer neurons. The values below the receptive field along the input tensor's full depth are transformed by a non-linear function. However, in contrast to fully connected layer neurons, the same set of weights (referred to as a kernel) is used for all receptive fields in the input volume resulting into a transformation that has a form of convolution across the input. A kernel is convolved across W and H spatial dimensions. Then, a different kernel is again convolved across the input volume producing another 2D tensor. Aligning up the output tensors into a $C \times W0 \times H0$ volume assembles the layer's output feature map. This is an important property of convolutional neural networks because each kernel detects a specific feature in the input. For example, in the first layer, the first kernel would detect presence of horizontal lines in the receptive fields, the second kernel would look for vertical lines, and similarly further on. In fact, learning such types of detectors in the bottom layers is typical for CNNs. The design of CNNs has an immensely practical implication – since a kernel is convolved

across the input utilizing the same set of weights and it covers only the receptive field, the number of parameters is significantly reduced. Therefore, convolutional layers are less costly in terms of memory usage and the training time is shorter.

Pooling Layer

Convolutional layers are designed in such a way the spatial dimensions are preserved and the depth is increased along the network flow. However, it is practical to reduce spatial dimensions, especially in higher layers. Dimensions reduction can be obtained by using stride when convolving, leading to dilution of receptive fields overlap. Nevertheless, a more straightforward technique was developed called a pooling layer Pooling Layer Convolutional layers are designed in such a way the spatial dimensions are preserved and the depth is increased along the network flow. However, it is practical to reduce spatial dimensions, especially in higher layers. Dimensions reduction can be obtained by using stride when convolving, leading to dilution of receptive fields overlap. Nevertheless, a more straightforward technique was developed called a pooling layer.

Image Classification

In the most recent work in computer vision, variations of neural networks trained with stochastic gradient descent are mostly used. These models are trained to classify an image, i.e. assign a class label to it . It proved to be very efficient to utilize a pre-trained image classification model in similar tasks. From the great variety of pre-trained models, let us mention VGG-16 and Inception V3 that set the state-of-the-art performance. Most recently, Inception V4 was introduced proposing a new framework of a convolutional layer that adds the feature map to its input and outputs this sum.

Object Detection

In the dense captioning task , an image is described with a set of statements, each corresponding to a salient region in the image. Thus, different phenomenons present in the image need to be detected and localized in order to describe them. The most advanced methods solving such a task (object detection) draw again on CNNs, such as YOLO or Fast R-CNN . Lately, Ren et al. focused on generating boxes with a fully convolutional network by converting anchors into region proposals (Faster R-CNN), yet they do not propose end-to-end trainable architecture. Drawing on Faster R-CNN, a model entirely trained with back-propagation has been proposed by Johnson et al., called FCLN . In terms of time performance, SSD by outperforms others.

Image Captioning with Convolutional Neural Networks

Recurrent Neural Networks

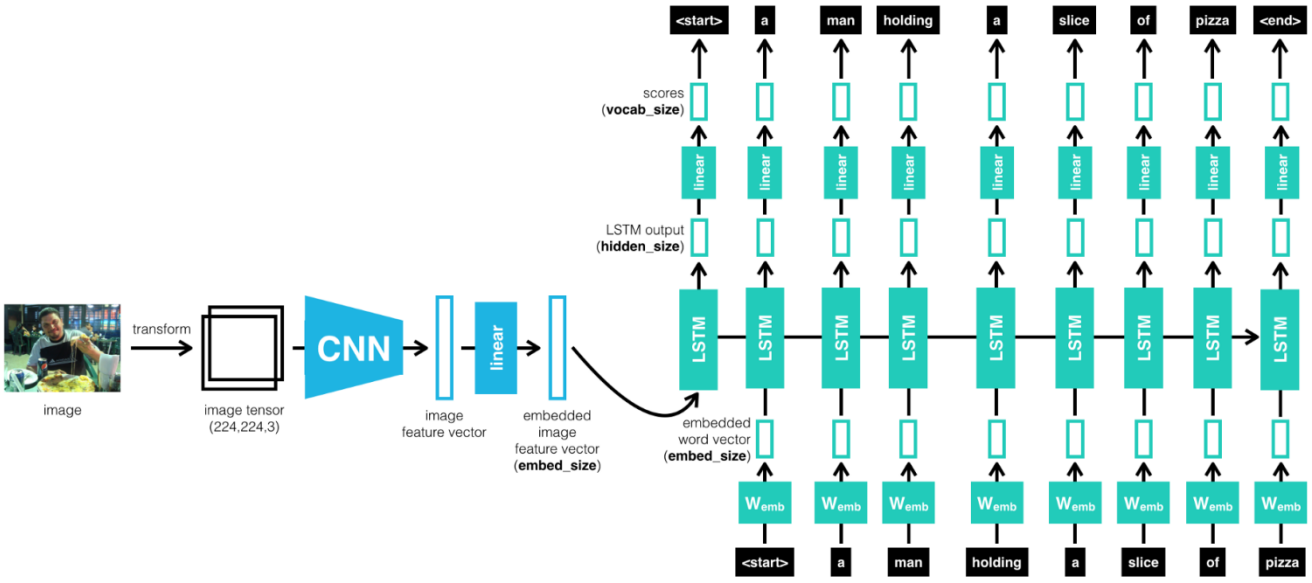
Assuming an image have been decoded into a feature map, recent publications adopt Recurrent Neural Networks (RNNs) to construct a language model that generates text. Ability of RNNs to preserve long-term relations and generate sequences was studied by Bahdanau et al. [Calling on these notions, a multimodal RNN was introduced . Further work, on the other hand, simplified the language model and developed other RNN designs. Karpathy and Fei-Fei exploited a Bidirectional RNN , whereas for example Kiros et al. rather drew on the LSTM recurrence by propagating the visual information from the decoder directly to the language model. Karpathy et al. showed why LSTM in particular provides surprisingly good results despite their complex architecture. A certainly interesting idea was introduced by Bengio et al. who took into account spatial attention – a map that highlights areas of an image that shall be described by the language model.

5 Experiments

Having stated the fundamentals of image captioning models, this section presents the experiments conducted to verify hypotheses regarding their performance. It is important to underline the fact that improvement is traceable as long as there is sufficiently defined evaluation function. In the problem of dense captioning, there are several aspects to be measured. Namely, a generated bounding box shall be positioned correctly and a caption shall be well-written and meaningful. Note that the later is defined vaguely and, thus, needs to be specified

Figure 1:

Image Captioning with Convolutional Neural Networks



A visualization of self-attention in our proposed Object Relation Transformer. The transparency of the detected object and its bounding box is proportional to the attention weight with respect to the chair outlined in red. Our model strongly correlates this chair with the companion chair to the left, the beach beneath them, and the umbrella above them, relationships displayed in the generated caption.

Overview of Object Relation Transformer architecture. The Bounding Box Relational Encoding diagram describes the changes made to the Transformer architecture In this work, we propose and demonstrate the use of object spatial relationship modeling for image captioning, specifically within the Transformer encoder-decoder architecture. This is achieved by incorporating the object relation module of [9] within the Transformer encoder. The contributions of this paper are as follows: • We introduce the Object Relation Transformer, an encoder-decoder architecture designed specifically for image captioning, that incorporates information about the spatial relationships between input detected objects through geometric attention. • We quantitatively demonstrate the usefulness of geometric attention through both baseline comparison and an ablation study on the MS-COCO dataset. • Lastly, we qualitatively show that geometric attention can result in improved captions that demonstrate enhanced spatial awareness.

Image Captioning with Convolutional Neural Networks

Dataset and Metrics

We trained and evaluated our algorithm on the Microsoft COCO (MS-COCO) 2014 Captions dataset . We report results on the Karpathy validation and test splits , which are commonly used in other image captioning publications. The dataset contains 113K training images with 5 human annotated captions for each image. The Karpathy test and validation sets contain 5K images each. We evaluate our models using the CIDEr-D , SPICE , BLEU , METEOR, and ROUGE-L metrics. While it has been shown experimentally that BLEU and ROUGE have lower correlation with human judgments than the other metrics , the common practice in the image captioning literature is to report all the aforementioned metrics.

Comparative Analysis

We compare our proposed algorithm against the best results from a single model1 of the self-critical sequence training (Att2all) the Bottom-up Top-down (Up-Down) baseline, and the three best to date image captioning models. Table 1 shows the metrics for the test split as reported by the

authors. Following the implementation of , we fine-tune our model using the self-critical training optimized for CIDEr-D score and apply beam search with beam size 5, achieving a 6.8% relative improvement over the Up-Down baseline, as well as the state-of-the-art for the captioning specific metrics CIDEr-D, SPICE, as well as METEOR, and BLEU-4.

Positional Encoding

Our proposed geometric attention can be seen as a replacement for the positional encoding of the original Transformer network. While objects do not have an inherent notion of order, there do exist some simpler analogues to positional encoding, such as ordering by object size, or left-to-right or top-to-bottom based on bounding box coordinates. We provide a comparison between our geometric attention and these object orderings in Table 2. For box size, we simply calculate the area of each bounding box and order from largest to smallest. For left-to-right we order bounding boxes according to the x-coordinate of their centroids. Analogous ordering is performed for top-to-bottom using the centroid y-coordinate. Based on the CIDEr-D scores shown, adding such an artificial ordering to the detected objects decreases the performance. We observed similar decreases in performance across all other metrics (SPICE, BLEU, METEOR and ROUGE-L).

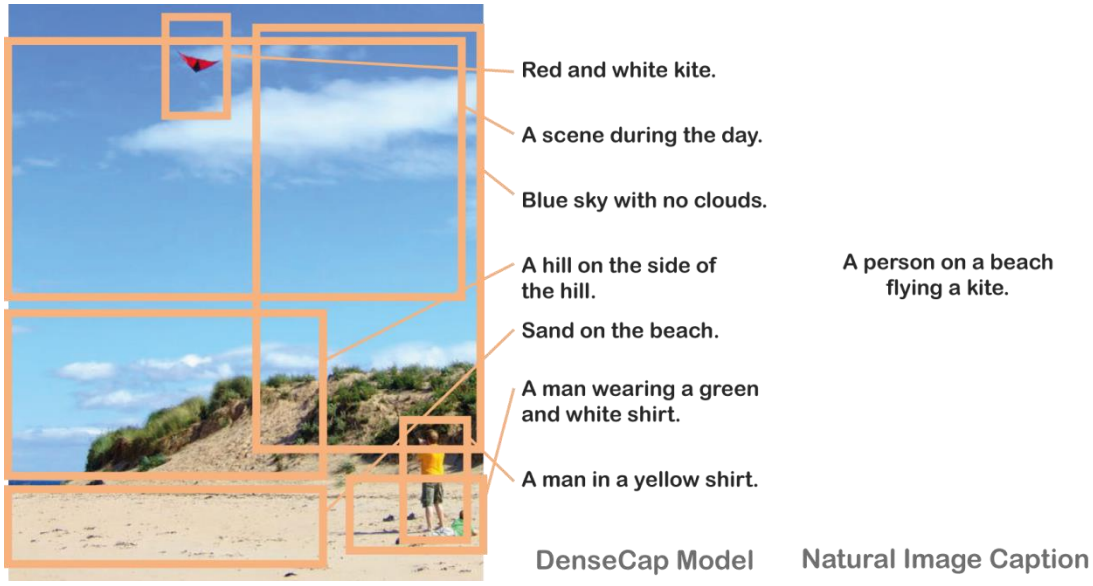


Image Captioning with Convolutional Neural Networks

Generated Captions

0.00946 A train on the tracks.
0.00000 Train on the tracks.
-0.00099 A man carrying a backpack.
-0.00099 A yellow line on the platform.
-0.00049 Trees behind the train.
-0.00049 Train tracks on the ground.
0.00000 People waiting for the train.
-0.00049 Trees growing on the side of the road.
0.00946 The train is yellow.
-0.00099 Front windshield of train.
0.00000 A clear blue sky.
0.00000 People walking on the platform.
-0.00149 A man wearing a black shirt.
-0.00049 Train tracks on the ground.
0.00000 A large green tree.
-0.00200 Person wearing black pants.
-0.00049 Lights on the train.
0.00946 People walking on the train platform.
-0.00099 A wooden door.
0.00000 Power lines above the train.
-0.00049 Green trees in the background.
-0.00149 Green bushes in the background.
-0.00049 Lights on the train.
-0.00049 Train tracks on the ground.
0.00000 A large tree in the distance.
0.00000 People waiting for train.
-0.00049 A platform on the platform.
0.00000 Gravel between the tracks.
-0.00049 Power lines on the pole.
-0.00049 Windows on the train.
0.00000 Train on the tracks.
-0.00049 A black bag on the ground.
-0.00049 Grass growing on the side of the road.
0.00000 Gravel between the tracks.
-0.00049 Train tracks on the ground.
-0.00049 Grass growing on the ground.
0.00000 Gravel on the ground.
-0.00049 Front of train is yellow.
-0.00099 Lights on the front of the train.
-0.00049 A black bag on the floor.
-0.00149 The woman is wearing black shoes.
-0.00049 A metal fence.
-0.00149 White clouds in blue sky.
-0.00099 Green grass on the ground.
0.00000 Power lines above the train.
-0.00099 A long white train.
-0.00049 A brown roof.
... (40 more)

Ground Truth Captions

A man awaits a train as it approaches.
A man waits patiently for the train to come.
A man waiting on a platform as a train approaches.
A train is approaching a platform with people.
A train is approaching the station.
A train is coming into a station.
A train is pulling in the station for the passengers.
A train pulls up to the station.
A man is waiting for the train.
A yellow train is on it's tracks.
Man waiting on the train.
People wait as a trolley approaches the station.
A man is waiting for the train to pass.
A train stopped at a train stop.
People waiting for the train.
A yellow train is at a station.
A man is standing on a platform near a train.
An incoming train arriving at a station.
A train pulling into a station.
Two people are waiting for a train to arrive.
(30 more)

Distinct Caption Set

Image Captioning with Convolutional Neural Networks

A young boy is playing basketball. 	Two dogs play in the grass. 	A dog swims in the water. 	A little girl in a pink shirt is swinging. 
A group of people walking down a street. 	A group of women dressed in formal attire. 	Two children play in the water. 	A dog jumps over a hurdle. 

Conclusion

We have presented the Object Relation Transformer, a modification of the conventional Transformer, specifically adapted to the task of image captioning. The proposed Transformer encodes 2D position and size relationships between detected objects in images, building upon the bottom-up and topdown image captioning approach. Our results on the MS-COCO dataset demonstrate that the Transformer does indeed benefit from incorporating spatial relationship information, most evidently when comparing the relevant sub-metrics of the SPICE captioning metric. We have also presented qualitative examples of how incorporating this information can yield captioning results demonstrating better spatial awareness. At present, our model only takes into account geometric information in the encoder phase. As a next step, we intend to incorporate geometric attention in our decoder cross-attention layers between objects and words. We aim to do this by explicitly associating decoded words with object bounding boxes. This should lead to additional performance gains as well as improved interpretability of the model.