Image Captioning using End-to-End Encoder-Decoder Paradigm

Yogesh Tripathi[CS16B044] and R. Pooja[CH15B056]

Indian Institute of Technology, Madras
Chennai-36

Abstract. In this project, we develop an end-to-end system using Encoder-Decoder framework to generate meaningful captions from raw image data. We use Convolutional Neural Network as encoder to perform feature extraction and a Recurrent Neural Network as decoder, a language model, which generates uses the semantic concepts captured by the CNN to generate a sentence which describes the contents of the image. Unlike the conventional approaches in Natural Language Generation, this approach does not split the tasks into stages like Text planning, Sentence planning and Linguistic realization, which in some sense, hard code the kind of sentences the system can generate. In this work, we just train a language model and the language model generates appropriate sentences.

Keywords: Natural Language Generation · Deep Learning · Convolutional Neural Network · Recurrent Neural Network.

1 Introduction

The objective of the image captioning task is to generate meaningful captions to images given as input. This task has a number of useful applications. This can improve accuracies of searches for images using their descriptions. The process of captioning photos and videos on social media can be automated. It can also be used to help visually challenged people get a better sense of their surroundings.

We implement image captioning by using an end-to-end encoder decoder framework. We use convolutional neural networks to perform the content selection task for us. An image is given as input to the CNN layer which encodes the information of the image and returns a global visual feature vector which captures the semantic features of the image and the relationships between them. This encoded information is then decoded using a recurrent neural network. A recurrent neural network is a sequential model which is used to handle sequential inputs like time-series data, text data and video processing where successive inputs are not independent. The CNN-RNN framework, once trained on image-caption data can then be used to generate descriptions for new images.

In the following sections we briefly explain the CNN and RNN framework. We also introduce the metric we used to evaluate the performance of our model

1.1 Convolutional Neural Networks

Image data can be represented as a multi-dimensional array of pixel intensities. An RGB image of pixel size 256 X 256 can be represented using an array of dimensions 256 X 256 X 3 (one 2D layer for every colour). In the convolution layer of the architecture, we have a number of filters convolve over the images to generate convolved features - 1 for each filter used, that in some sense detect and capture latent information and features from the images. These filters can be learnt through parameter estimation while training the model. A pooling layer can be added to the architecture in order to decrease the dimensionality, thereby reducing the number of parameters to be learned. This also helps minimize overfitting. Two popular types of pooling strategies include: max pooling and average pooling. We can add a number of convolution and pooling layers, based on the requirement of the task. Post this, in the case of classification tasks we flatten the matrices, add as many dense neural network layers as needed and then add the final dense layer which would be an applicable output layer.

We used the GoogleNet CNN architecture also known as the inception module to encode the images as it performed better than other architectures. In inception models instead of using filters of fixed dimensions which might not be very effective when we have principal components of images occupying areas of different proportions in the entire image, we use filters of different sizes that operate on the same level and then concatenate their outputs before sending them to the next layers. The number of parameters that need to be estimated can be reduced by factorizing the convolutions, for eg: use 2 layers of 3X3 filters (18 parameters) instead of 1 layer of 5X5 filter (25 parameters). This way the computational efficiency can also be significantly improved as convolving with a 5X5 filter, takes much longer than convolving with a 3X3 filter.

We also implemented the attention mechanism that gives us a sense of which region of the image to focus on. In this method, CNN is made to capture a set of vectors to represent different subregions of the image in addition to the global visual vector. This vector is also fed into the RNN which uses this additional information - the notion of space to determine the likelihood of the next word

1.2 Recurrent Neural Networks

Conventional models developed to handle independent data cannot be used on sequential data where the inputs depend on each other. Word order is paramount to understanding semantics of sentences and this can potentially get lost when text data is modelled using algorithms that disregard dependence between inputs. This is where sequential models like RNNs come in handy. These models account for the relations between the inputs. The number of inputs also need not be fixed in RNNs. The output at successive time-stamps(inputs) depends on both the current input and also the inputs given to the system in the past. In conventional RNNs as a sequence of inputs are fed in the identity of the initial inputs is completely morphed and it is impossible to recover the previous inputs from the current state. When analysing sequential data, we need mechanisms

that can forget redundant information, selectively store vital information and selectively transfer the relevant information. Long-short term memory(LSTM) and Gated Recurrent Networks are some models that possess these abilities. In our project we used Gated Recurrent Network with attention mechanism to decode the global feature vector and train them to caption images. GRU uses what can be understood as an update gate and a reset gate that effectively retain vital historic information without morphing them and are capable of ignoring irrelevant information.

1.3 Evaluation - BLEU (Bilingual Evaluation Understudy)

We used the BLEU metric to evaluate the performance of our model, i.e. to compare the machine generated captions with the captions in our dataset. This score is determined by computing the percentage of n-grams that were accurately retrieved. Precision is the fraction of retrieved result that is relevant. BLEU score is a modified form of precision, where the maximum number of occurrence of each n-gram in the retrieved result is clipped to be the maximum number of occurrences of that n-gram in the reference. This is done to penalize trivial models.

2 Related Work

In the recent past a number of deep neural end-to-end frameworks have been proposed to caption images. Researchers have drawn inspiration from models that perform well on machine translation and implemented them for generating image descriptions because the two fields are very similar. Deep CNNs are used to generate features from images, the output of the last activation layer of CNN is taken to represent the image features and subsequently fed into the RNN decoder along with the word sequences. During the training of parameters, a distribution of all the words in the vocabulary is computed and the occurrence of the true next work is maximized. A number of different combinations of architectures have been explored. Inspired by success in using the attention mechanism in sequence learning for machine translation, visual attention is being implemented for image captioning.

3 Experiments

The end-to-end encoder-decoder paradigm was implemented using pre-trained GoogLeNet network. The GoogLeNet network used was pre-trained on ImageNet data set. The output of final convolution layer from GoogLeNet was taken and flattened before obtaining an embedding of salient features in the image. The size of this embedding was fixed to be 256 throughout. The decoder was implemented using a Recurrent Neural Network using Gated Recurrent Unit(GRU) cells. The benefit with GRU cells over LSTMs is that there are lesser parameters than LSTM as forget gate parameters are removed. We augmented Attention

mechanism in the system so that the system captures more salient features for captioning based on attention values. Cross-entropy loss between one-hot representation of word in the caption and predicted word was used. The entire system was trained using backpropagation through time and Adam learning algorithm was used for learning. Flickr-30k data set was used in the experiment in which where 80% of data was used for training and 20% was used for testing. For evaluation of the system, BLEU score of the system was calculated over all examples and averaged to get the final BLEU score. Size of training set was varied from 10k to 30k to observe the performance increase with more training data. This is summarized in the Table 1 below. The experiments and design selected for this project very much rebuild the implementation of image captioning system in Show, Attend and Tell: Neural Image Caption Generation with Visual Attention by Kelvin Xu et al.[3]. Some of the captions generated by the system are shown on Page 5.

Training data size	BLEU-1 score	BLEU-2 score
10000	0.4322	0.2761
20000	0.4893	0.2903
30000	0.5157	0.3518

Table 1. BLEU score on varying size of training data

References

1. Xiaodong, HeLi Deng *Deep Learning for Image-to-Text Generation: A Technical Overview*
2. Daouda Sow,Zengchang Qin *A Sequential guiding network with attention for image captioning*
3. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*

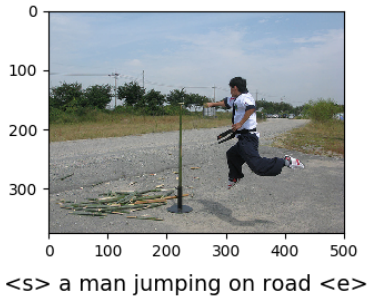

Fig. 1. Instance 1

Fig. 2. Instance 2

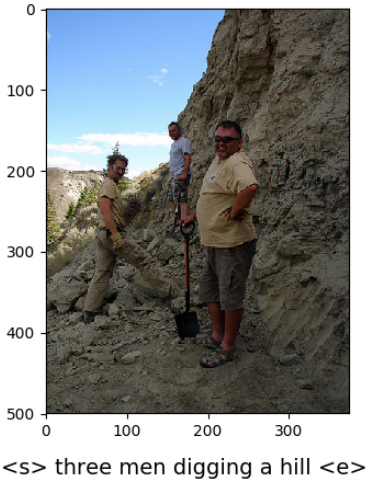

Fig. 3. Instance 3

Fig. 4. Instance 4

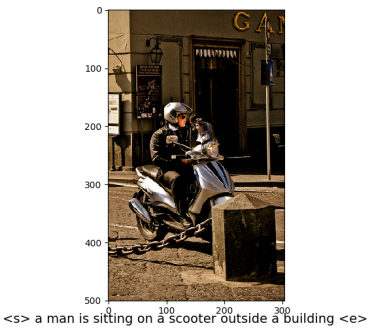

Fig. 5. Instance 5

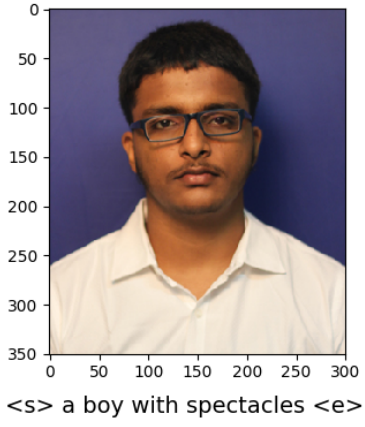

Fig. 6. Instance 6