Is there any practical approach to generate automated stories using visual data?

Arpit Garg
Faculty of ECMS
The University of Adelaide
Adelaide SA Australia
arpit.garg@adelaide.edu.au

What is your research question?

I plan to focus on the images to generate stories using artificial intelligence. In this method, we are mainly focused on the sequence to sequence model to create stories automatic with the use of image sequence encoder, which is also dependent on the previous sentence. The research question will be: Is there any approach that can be followed to reach state of the art for this process?

Why do you think the question is important?

There was not much work has been done on generating stories using images and giving the best approaches for these purposes. This question is essential as it will help to increase the software quality of the social platforms by reducing the risks of sharing fake and wrong news. If anyone can answer this question, then this could work as the amplification or verification system for the various online platforms. This research will not only help to improve the software quality but also increases developer productivity and reduces human efforts.

What does previous research tell you about your question?

I have looked for different researches and combined them all in this research. Previous related work is divided into three different categories: 1. Describing the images in isolation. 2. Describing the images in sequence. 3. Images in series for stories.

For describing the images in isolation. Work done by Karpathy[1] gave the features extracted from the convolutional neural network (CNN) and recurrent neural network (RNN) as the input and found that the answer to this
question is same as image captioning. They have also reached state of the art for image captioning. In Describing the images in sequence, Yao[2] tried to extract the features from images using 3D CNN in series and then used LSTM (Long Short-Term Memory) as the global approach for the structure. They have used the pre-trained CNN model for extraction of images and used RNN for temporal behaviour. In the latter approach, Liu used the concept of Bidirectional Attention-based Recurrent Neural Network (BARNN) model, which will generate the relationship between the sentences according to images in sequence. While using this process, Liu[3] also use three related RNN model to encode all the images.

It can be observed from all the previous results that there are different approaches available to answer this question, but none of the approaches is successful. Now we need to select the most effective method with some modifications in them based on different factors and parameters.

What data will you need to answer the question?

The primary source of data required to answer this question is Visual Storytelling Dataset (VIST). Training and testing of the model can be done on this dataset. This dataset consists of 50,000 unique stories and 20,211 photos. The flick is the leading source for the collection of images using its API. This album included a minimum of 10 images and maximum 50 images, and all these images were taken at an interval of 48-h. This provides the relationship between images and stories. Amazon Mechanical Turk workers were asked to randomly select five albums and write a story related to each album. Every image is associated with some story, and there is a minimum five-line story present with every image. The main thing to focus on the dataset is the gender replaced all the names according to the image.

This dataset is divided into two parts description of images in isolation and stories of images in sequence. The only problem with the VIST dataset that I can figure out is that all these stories were created by people only then they are processes so there might be the possibility that there is no sequence or the flow of the story. There is a high chance that in some cases, the story is not correlated with the image sequence. But we are not considering this case for this project. Therefore, we are not expecting perfect stories as a result. We are only concern that our system will be able to generate stories in our language. We only want our system at least contains the word related to the image sequence. For all these purposes we are using VIST and Flick API for our dataset.

How are you going to analyse this data?

We start the analysis of data by dividing it into three parts. 80% of the data will be used for training the model and 10% for cross-validation and 10% for the testing purposes. First, the extraction of features from all the images and store all these in files is required, and all this can be done using Alexnet. The file is stored in the format of Hierarchical Data Format version 5 (HDF5). After this, each features vector obtained need to be associated with the vectorised sentence. We will use the dataloader to align each feature with the respective sentence.

We can use Attention-based Neural Translation Machine mechanism to train an encoder and decoder. It directly converts the source sentence to target sentence by modelling the conditional probability p(y|x) and divide the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j|y_{<j}, \boldsymbol{s}\right)$$

RNN will be used as a hidden unit for the decoder and CNN will be used for the encoder. Similarly, RNN with LSTM can be used both as encoder and decoder. We can parameterise the probability by :

$$p\left(y_j|y_{<j}, \boldsymbol{s}\right) = \text{softmax}\left(g\left(\boldsymbol{h}_j\right)\right)$$

where,

$$\boldsymbol{h}_j = f(\boldsymbol{h}_{j-1}, \boldsymbol{s})$$

After using LSTM, our objective for training is

$$J_t = \sum\nolimits_{(x,y)\in\mathbb{D}} -\log p(y|x)$$

Using this approach by considering the previous sentence, the current sentence is generated with the use of encoder and decoder. We will try to follow the sequence to sequence approach. In this, our encoder is divided into two parts. The first part is related to the behaviour of the image sequence, and the second encoder is dealing with modelling of the sentence with the previously generated sentence. RNN can learn temporal dependencies, so it is used here for sequence modelling. The encoder that is dealing with the behaviour aligns every sentence with the corresponding images. The decoder module is also RNN which reproduce all the information encoded.

In the training process, we will give the encoder image sequences and previous sentences, and in the decoder, we will provide current-sentences. For doing all these steps, we also converted our images from RGB to BGR and cropped them even as we need to give these as the input to AlexNet. The words we are using would be the most frequent words that will be appearing in the stories. Based on the common observation, we have decided that our length of each sentence will be approximately equal to 20 words per sentence. Categorical cross-entropy will be used as the loss function, and the learning rate is set to 0.0001 based upon the improvement.

In contrast, training and ADAM algorithm will be used for the optimisation. To measure how close the generated translation is, we used the METEOR and BLEU scores as they are used when we are considering Neural Machine Translation. They measure between 0 to 100 to show how close the results are. After training our model, we will just try our model with cross-validation and testing data. By analysing the data in this manner, we can comment on the question if the approach is valid or not.

How does this relate to the course material?

In terms of research, this question is a critical type research question as the knowledge of the study is objective and prefer quantitative approaches so it can be referred to as positivist knowledge. By this, I am aiming to verify my question and find the most effective method based on quantitative analysis. The study is eclectic as this question is more focused towards the problem and the problem is that we need to find if there is any effective approach to generate stories automatically using images sequence and to answer this we have used various models and methodologies. It is more kind of Statistical based approach where we have limited data collection, and we are interpreting based on that. The course material also focuses on research ethics. In this research there all the data is collected from VIST, which is open to all. It's main aim to reduce the time on pre-processing so that everyone can focus on approach. So, no anonymous data have been referred to while answering this question. The main purpose of the course material is to come with the solutions for the general problems related to software and programs. In this proposal, we try to find the appropriate approach to answer these questions.

REFERENCES:

[1] Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. CoRR abs/1412.2306 (2014)

[2] Yao: Describing videos by exploiting temporal structure (2015)

[3] Liu, Y., Fu, J., Mei, T., Chen, C.W.: Let your photos talk: generating narrative paragraph for photostream via bidirectional attention recurrent neural networks. AAAI Conference on Artificial Intelligence, February 2017