ЖАРАТЫЛЫСТАНУ ЖӘНЕ ТЕХНИКАЛЫҚ ҒЫЛЫМДАР

SCIENCE AND TECHNOLOGY

IRSTI 06.58.45

B. Kynabay[1], A. Aldabergen[2], S. Shamiluulu[3]
[1,2,3]Suleyman Demirel University, Kaskelen, Kazakhstan
[3]Yahoo company

DEVELOPMENT OF IMAGE CAPTIONING MODEL BY USING DEEP LEARNING TOOLS

Abstract. Automatic formation of image description is one of the most challenging and popular topics in the field of deep learning (DL) for nowadays. In this work image captioning convolutional neural network model (VGG16) is developed by implementing deep learning tools and techniques. Specifically, this work's resulting product can be implemented in a system that answers to the question, based on the image given. Image captioning includes two main sub-processes: image processing and natural language processing. The model was constructed from 16 layers and it uses Flicker 8K data-set of images for captioning. The model was evaluated by using BLEU metric and its value was nearly 0.75, it takes one image as an input and returns one description for that image.

Keywords: Deep learning, convolutional neural network, image captioning, VGG16, visual question answering, bilingual evaluation understudy.

***

Аңдатпа. Сурет сипаттамасын автоматты түрде қалыптастыру - қазіргі кездегі терең оқыту саласындағы ең күрделі және танымал тақырыптардың бірі. Бұл жұмыста терең оқыту құралдары мен әдістерін қолдану арқылы консолюциялық нейрондық желі үлгісі (VGG16) жасалынған. Атап айтқанда, осы жұмыстың нәтижесі берілген суретке негізделген сұраққа жауап беретін жүйеде жүзеге асырылады. Сурет

сипаттамасын жасау үрдісіне екі негізгі ішкі үрдіс кіреді: суретті өңдеу және табиғи тілді өңдеу. Үлгі 16 қабаттан тұрғызылған және суретке арналған сипаттамалар ретінде Flicker 8K мәліметтер жиынтығын пайдаланады. Үлгі BLEU метрикасын қолдану арқылы бағаланды және оның мәні шамамен 0,75 болды, ол енгізілетін ақпарат ретінде бір сурет қабылдайды және сол суреттің сипаттамасын береді.

Түйін сөздер: Терең оқыту, конвульсиялық нейрондық желі, сурет сипаттамасын жасау, VGG16, сұраққа визуалды жауап беру, екі тілді бағалау.

***

Аннотация. Автоматическое формирование описания изображения является одной из самых сложных и популярных тем в области глубокого обучения на сегодняшний день. Модель нейронной сети (VGG16) разработана для создания описания изображения с использованием инструментов и методов глубокого обучения. В частности, полученный продукт этой работы может быть реализован в системе, которая отвечает на различные вопросы, основываясь на представленном изображении. Создание описания изображений включает два основных подпроцесса: обработка изображений и обработка естественного языка. Модель была построена из 16 слоев и использует описания Flicker 8K для изображений. Модель была оценена с использованием метрики BLEUю Точность модели около 0,75, она принимает одно изображение в качестве входного данного и возвращает одно описание для этого изображения.

Ключевые слова: Глубокое обучение, сверточная нейронная сеть, описание изображения, VGG16, визуальный ответ на вопросы, двуязычная оценка.

Introduction

In the past several decades, deep learning, which can be described as a class of machine learning algorithms, has showed noticeably good results across a variety of problems. In one word it is all about accuracy. Nowadays, recognition accuracy of deep learning have reached the highest plane than ever before. For instance, on a daily life DL can help electronics satisfy user demands in much more better level, it is used in automation of processes in almost every field of life. In one of the recent articles about DL it is stated that latest approaches in DL has reached the point where it outperforms humans, in some tasks like object classification [1]. Deep learning has recently arose as a highly successful paradigm for big data. By the means of technological

advances deep learning has given noticeable contributions in several basic artificial intelligence tasks in the fields of image processing and computer vision.

Image is a depiction of visual perception or object in a binary form that carries very large amount of information. It will be very useful for different kind of real-life tasks if machines are able to parse all the information from any image. However, for computers, to parse all that information is not as easy as for humans, but possible.

Main purpose of this work is to develop a model that can perform this task as accurate as possible. If normally humans use their languages for image captioning process, machines are supposed to use it also. Model that can construct the description of an image can be built by the help of natural language processing techniques. Most of the previous works that are done on this task were not very accurate as they used to deal with exact defined syntax (sentences) and hard coded features [2]. The reason for that is the model which was built on exact descriptions for exact images. In order to overcome such problem it is necessary to make the model independent of such hard coded descriptions.

In this paper, pre-trained VGG16 model was used for image captioning process. The data used have over 8000 instances with their 5 different descriptions. Several research works that are done on this topic are described in the section II. Proposed methods, used materials and achieved results are explained in details in sections III and IV. The last section contains discussion, conclusion and further work on this problem.

Literature review

Indeed there are a lot of works done on this topic as intersection of vision and language are now one of the significant questions of modern research. As computers do not have intelligent reasoning, detecting objects on the image and determining their relationship by constructing the context of the image is challenging task for them.

There are two most common ways of doing image captioning: generative and retrieval based methods. As an example of retrieval based method we can take Im2Txt model [3]. This model was constructed by Vicente Ordonez, Tamara L Berg and Girish Kulkarni. General system is contained of two parts: matching the image and generation of the caption. When an image is given to the system it will be matched with the image and its caption which is in the database. After the images are matched, received objects of high level will be compared with original input data. One of the disadvantages of this model is being overfitted, it can construct captions only from already prepared ones in the dataset. However, this problem is solved in generative models.

Because, generative models let to construct new sentences (descriptions) for the model. Model proposed in the work of Vinyals et al. called “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge” is one of the examples of generative type of models. This model implements neural machine translation and newest image recognition techniques, by using LSTM and model of inception-v3 [4].

If to look at researches on the task of visual question answering (where the main task is image captioning) they range from constrained settings to freeform natural language questions and answers. For example, the authors of research work “A visual Turing test for computer vision systems” propose a system to generate binary questions from templates using a fixed vocabulary of objects, attributes, and relationships between objects [5]. They constructed a ”visual Turing test”: an operator-assisted device that produces a stochastic sequence of binary questions from a given test image. Engine will construct some question, afterwards question will be checked by the operator and either the answer will be given or will state the question was ambiguous. Another research work called “Video And Text Parsing for Understanding Events and Answering Queries” propose a multimedia analysis framework to process video and text jointly for understanding events and answering user queries [6].

Materials and methods

a. dataset

Flicker 8K dataset is a collection of 8000 images. Each image has 5 different captions that describe events and objects on the image. Images were collected from 6 various Flickr groups. It does not contain celebrities or well-known places, images were filtered manually to have different kinds of situations and scenes [7]. General overview of the dataset can be seen on the Fig. 1.

Materials used in this work are: VGG16 model (which is pre-trained by ImageNet data of images and has 1000 different classes) and Flicker 8k dataset which will be used for training the model with its captions.

For code compilation free Jupyter environment - Colaboratory was used. It provides with high RAM and has larger speech compared to the ordinary laptop speed. The code is written in Python 3.5 language.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog are fighting

young girl with pigtails painting outside in the grass .
there is a girl with pigtails sitting in front of a rainbow painting .
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
a little girl is sitting in front of a large painted rainbow .
a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

Figure 1. Flicker 8k dataset

b. methods: CNN model

Neural network architectural image captioning model is implemented in keras framework. Keras is one of the simplest, fastest and mostly used human oriented APIs. The VGG16 model imported from keras has 16 layers and over 138 000 000 parameters (Refer to Fig. 2).

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000

Total params: 138,357,544
Trainable params: 138,357,544
Non-trainable params: 0

Figure 2. CNN model summary

General methodology consists of main five parts:

- Text and image preparation
- Linking the captions and images
- Model construction and training
- Prediction
- Evaluation

General workflow of the process is described in the diagram below (Refer to Fig. 3).

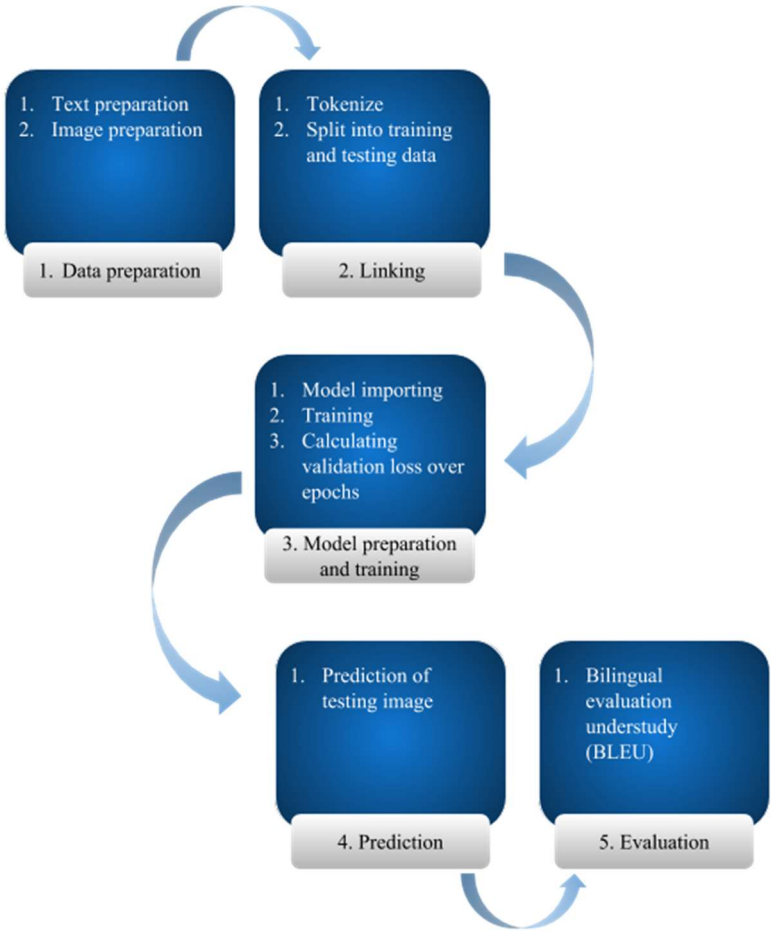

Figure 3.Process flow diagram

Results and discussion

a. text and image preparation

First of all distribution of the words in the text was visualized in order to prepare more accurate and informative text for captioning. It shows frequency of each occurring word. Afterwards, some articles like "a", "the" or punctuation symbols were removed, as they did not affect the meaning of the sentence (Refer to Fig. 4 and Fig. 5).

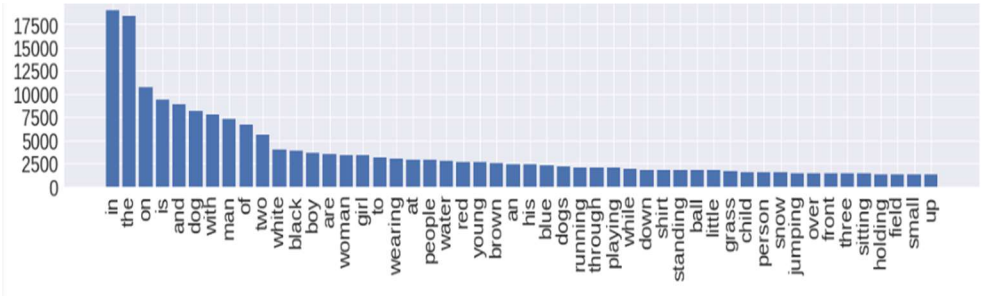

Figure 4. Initial word dataset distribution

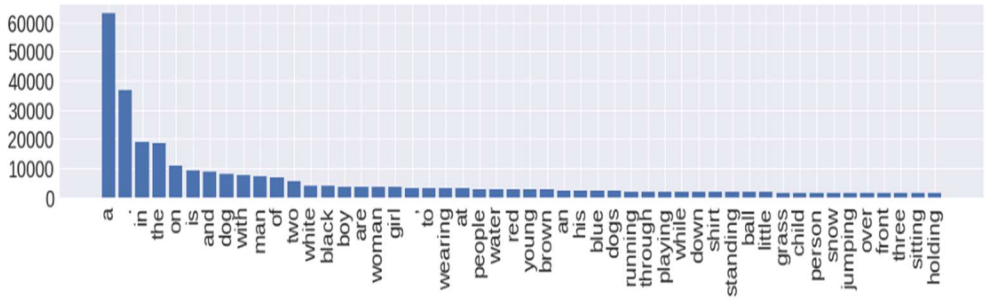

Figure 5. New word dataset distribution

Also for each caption start and end tokens were added, in order to easily identify the start and the end of the sentence. Ready weights for the pre-trained model VGG16 was downloaded from the GitHub repository. This model is able to classify images into 1000 classes.

b. linking the captions and images

In this part all the images were linked with their descriptions data were divided into training and testing sets (Refer to Fig. 6).

```
# captions/images = 3600
 (36714, 30) (36714, 4096) (36714, 3786)
# captions/images = 1200
 (12173, 30) (12173, 4096) (12173, 3786)
```

Figure 6. Data division into training and testing sets

c. model construction and training

The model is 16 layers deep and has over 138 mln of parameters (Refer to Fig. 1). It was trained about 12 minutes and went through 5 epochs (Refer to Fig. 7).

```
Train on 36714 samples, validate on 12173
Epoch 1/5
 - 152s - loss: 5.4281 - val_loss: 4.9744
Epoch 2/5
 - 151s - loss: 4.4984 - val_loss: 4.6684
Epoch 3/5
 - 151s - loss: 4.0013 - val_loss: 4.5030
Epoch 4/5
 - 151s - loss: 3.6526 - val_loss: 4.5109
Epoch 5/5
 - 151s - loss: 3.3584 - val_loss: 4.5710
TIME TOOK 12.63MIN
```

Figure 7. Model training

d. prediction

This step of methodology is about testing the constructed ready model. Randomly chosen images are given as an input one by one and the model returns one description for each. Some captions were not right and some were right (Refer to Fig. 8).

e. evaluating

There is a metrics called Bilingual Evaluation Understudy (BLEU) that is used for measuring the likeliness of one sentence with other given sentences. Its value varies between 0 and 1. BLEU is very well-known and easily calculated accuracy metric. While evaluating the constructed model BLEU values were good enough, it has shown result over 0.75, which means a good captioning (Refer to Fig. 9).

Figure 8. Predicted captions

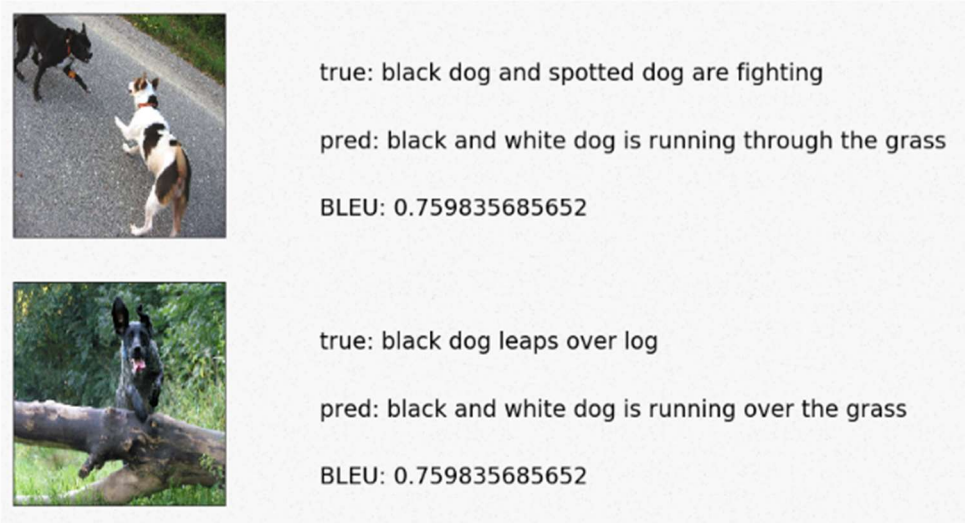

Figure 9. BLEU accuracy scores

Conclusion

Results have shown that it is possible to build image captioning system by the help of DL tools and techniques. The predicting model was constructed from VGG16 well trained and prepared neural network model. Model has shown an accuracy about 0.75. However, in some cases model have shown very irrelevant descriptions also, it can be caused by the size of the data given for training the model, because 7000 is not a big amount for model training in images classification problems. This limitation was caused by the lack of RAM memory on the laptop, only about 7000 images could be processed and sent for training the model. In further works, it will be more appropriate to connect additional GPU, or to buy a server in some large servers like Amazon or to process the images in a PC with larger amount of memory.

Acknowledgement

Thanks for the dataset provided by *Flicker groups* and *Colaboratory environment* for providing with additional memory and speed.

References

1 What Is Deep Learning? MATLAB for Artificial Intelligence, [electronic resource]// URL: What Is Deep Learning? MATLAB for Artificial Intelligence web., (accessed: April 2019).

2 Shah, P., Bakrola, V., Pati, S. Image Captioning using Deep Neural Architectures, *2017 International Conference on Innovations in*

information Embedded and Communication Systems (ICIIECS). IEEE, pp. 1-4.
3 Ordonez, V., Kulkarni, G., Berg, T.L. Im2text: Describing images using 1 million captioned photographs, in Advances in Neural Information Processing Systems, 2011. pp. 1143-1151.
4 Vinyals, O.A., et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 39 *(*4), (2016): pp. 652-663.
5 Geman, D., et al. A visual Turing test for computer vision systems, *Proceedings of the National Academy of Sciences*, 112 (12), (2015): pp. 3618-3623.
6 Tu, K., Meng, M., et al. Joint Video and Text Parsing for understanding Events and Answering Queries, *IEEE MultiMedia*, 21 (2), (2014): pp. 42-70.
7 Flickr8k Data, [electronic resource]// URL: https://forms.illinois.edu/sec/1713398. (accessed: June 2019).