International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Image Captioning: Transforming Objects into Words

Nafees Shaikh[*1], *Divyam Singh Thakur*[*2], *Gajendra Goswami* [*3], *Ajay Khatri*[*4]

*1Student, Department of Computer Science Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.
*2Student, Department of Computer Science Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.
*3Student, Department of Computer Science Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India.
*4Assistant Professor, Department of Computer Science Engineering, Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India
nafeesshaikh492@gmail.com,divyamthakur222@gmail.com , suyashgoswami2014@gmail.com, ajaykhatri@acropolis.in

Abstract

Scene interpretation, medical image analysis, robotic perception, video surveillance, augmented reality, and picture compression are just a few of the uses of image segmentation in image processing and computer vision. Various picture segmentation algorithms have been developed in the literature. Due to the success of deep learning models in a variety of vision applications, there has recently been a significant amount of research directed towards developing picture segmentation algorithms utilizing deep learning models. We present a comprehensive review of the literature at the time of writing, covering a wide range of groundbreaking works for semantic and instance-level segmentation, including fully convolutional pixel-labeling networks, encoder-decoder architectures, multi-scale and pyramid-based approaches, recurrent networks, visual attention models, and generative models in adversarial settings. We look at the similarities, strengths, and problems of these deep learning models, as well as the most extensively used datasets, describe results, and identify possible future research avenues in this field..

Key-Words: - Image Segmentation, Machine Learning Algorithms,Deep Learning Models

I. Introduction

Automatically describing the content of an image using properly formed English words is a difficult task, but it could have a significant impact, such as assisting visually challenged persons in better understanding the content of web images. For example, this task is much more difficult than the well-studied image classification or object recognition tasks that have been a major focus in the computer vision community [27]. Indeed, a description must communicate not just the objects represented in an image, but also how they connect to one another, as well as their traits and the activities they engage in. Furthermore, the aforesaid semantic knowledge must be communicated in a natural language such as English, necessitating the use of a language model in addition to visual comprehension.

Most traditional image captioning systems use an encoder-decoder structure, in which an input image is encoded into an intermediate representation of the information contained within the image, and then decoded into a descriptive text sequence, which is inspired by neural machine translation. This encoding can be made up of a single feature vector generated from a CNN or numerous visual features extracted from different parts of the image. The zones can be evenly sampled in the latter scenario, or directed by an object detector, which has been found to improve performance. While these detection-based encoders are cutting-edge, they currently do not take into account spatial relationships between observed items, such as relative location and size. This information, on the other hand, is frequently necessary for comprehending the substance of an image and is used by humans when reasoning about the physical world. For example, relative position can help identify "a female riding a horse" from "a girl standing by a horse." Similar to how relative size can help distinguish between "a woman playing the guitar" and "a woman playing the ukelele," relative size can help distinguish between "a woman playing the guitar" and "a woman playing the ukelele." Including spatial relationships in object detection has been shown to boost performance, as illustrated in. Furthermore, positional correlations are frequently recorded in machine translation encoders, particularly in the Transformer, an attention-based encoder design. As shown in Figure 1, the usage of relative positions and sizes of detected items should assist picture captioning visual encoders as well.

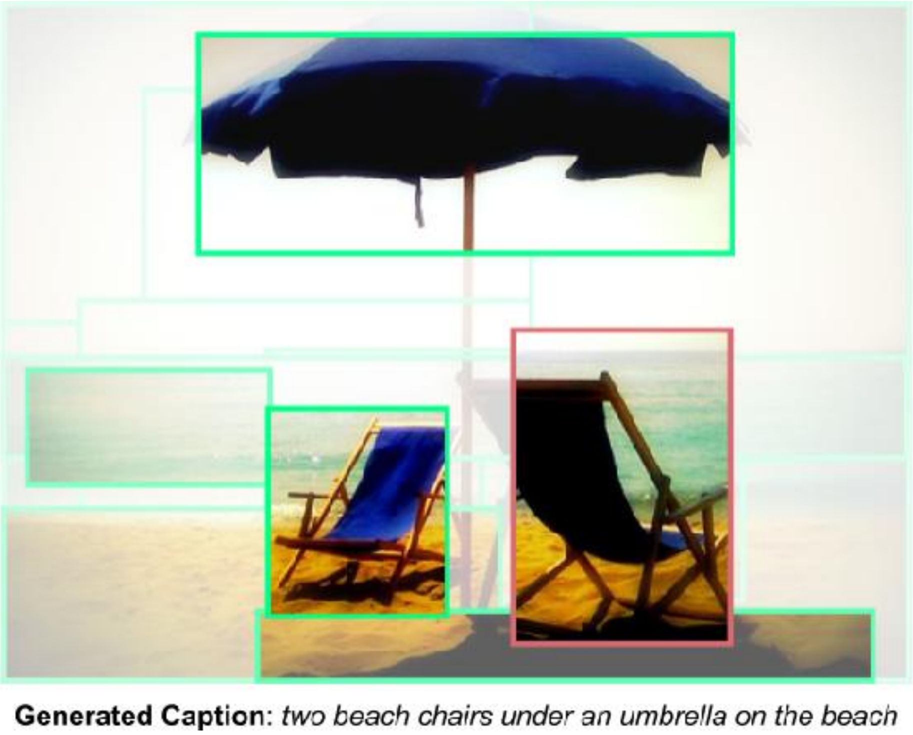

Generated Caption: *two beach chairs under an umbrella on the beach*

Figure 1 shows how our suggested Object Relation Transformer visualizes self-attention.

The attention weight with relation to the red-outlined chair determines the transparency of the identified object and its bounding box. Our model shows a strong association between this chair and the companion chair to the left, the beach beneath them, and the umbrella above them in the generated caption.

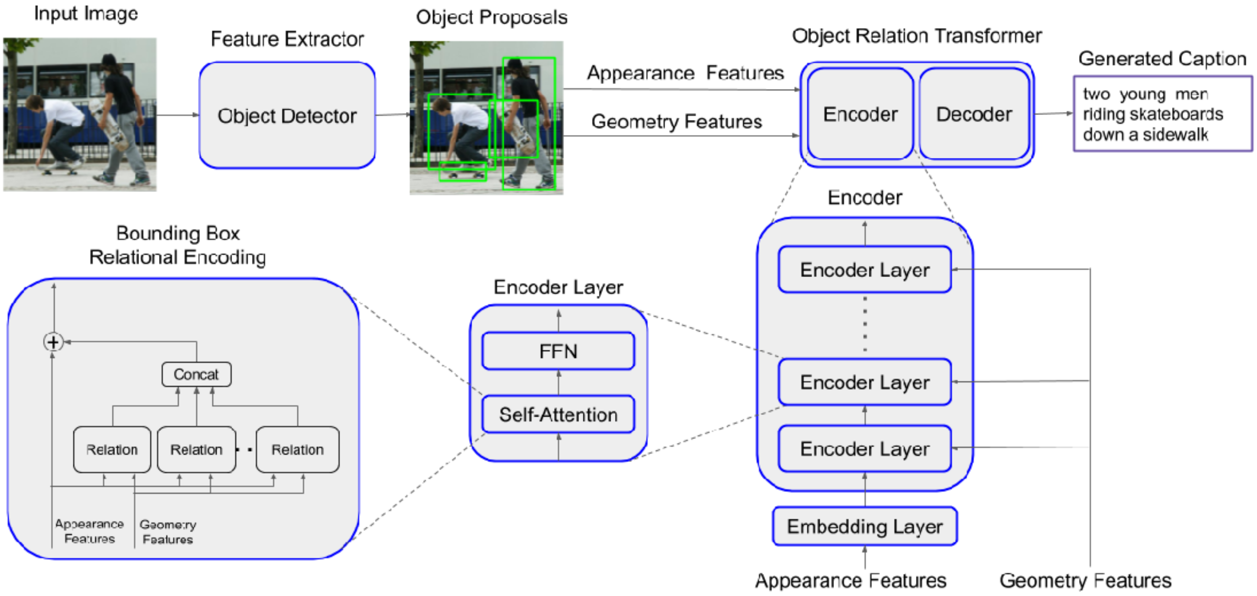


Figure 2: Overview of Object Relation Transformer architecture. The Bounding Box Relational Encoding diagram describes the changes made to the Transformer architecture

The use of object spatial relationship modeling for picture captioning is proposed and shown in this paper, specifically inside the Transformer encoder-decoder architecture. This is accomplished by adding the Transformer encoder's object relation module from [9]. The following are the contributions of this paper: • We present the Object Relation Transformer, an encoder-decoder architecture created exclusively for picture captioning that uses geometric attention to incorporate information about the spatial connections between input recognized objects. • Using the MS-COCO dataset, we quantitatively illustrate the utility of geometric attention through baseline comparison and an ablation analysis. • Finally, we show that geometric attention can lead to better captions that reveal increased spatial awareness.

II. Related Work

We provide background information on recurrent neural networks and image caption synthesis in this section. Several methods for automatic image caption generation have recently been tested. The researchers first proposed employing a graphical model based on human-engineered features to learn a mapping between images, meanings, and captions. The multi-model pipeline in suggested the pioneering use of neural networks for image caption generation, demonstrating that neural networks could decode image representations from a CNN encoder and that the resulting hidden dimensions and word embeddings contained semantic meaning (i.e. "image of a blue car" - "blue" + "red" produces vectors similar to "image of a red car"). Approaches from the top down: Following these early attempts, and demonstrated the use of these models on video captioning tasks, using more recent CNNs for encoding and replacing feedforward networks with recurrent neural networks. In particular, LSTMS proved the application of these models on video captioning tasks. Unlike prior work by, it demonstrated that an LSTM that did not receive the image vector representation at each time step could nevertheless deliver state-of-the-art results. These papers all had one thing in common: they represented images as the top layer of a massive CNN (thus the name "top-down" because no individual objects are identified) and built end-to-end trainable models. Bottom-up approaches: instead of approaching the problem as one large problem, divide it into two smaller ones. To begin, they train a CNN and a bi-directional RNN to learn to map images and caption fragments to the same multimodal embedding, achieving state-of-the-art performance on information retrieval tasks. Second, they train an RNN to construct a caption by combining the inputs from numerous object fragments recognized in the original image. Rather than working on a single image representation, this built on past work by allowing the model to collect information about specific items in the image. A similar line of research was explored in which object detectors were trained to identify image fragments and a three-step pipeline was devised for integrating two of these discovered fragments into a caption. However, one disadvantage of these models is that they could not be trained from beginning to end.

Recent advances in NLP, including the Transformer architecture [23], have resulted in significant performance gains for tasks like translation [23], text generation [4], and language understanding [19]. The Transformer was used to perform image captioning in [22]. By partitioning the image into 8x8 segments, the authors investigated extracting a single global image feature as well as evenly sampling features. In the latter example, the feature vectors were given to the Transformer encoder in a sequential order. In this study, we suggest using the bottom-up approach of [2] to improve on this uniform sampling. Because, unlike an RNN, the Transformer design has no idea of order for its inputs, it is particularly well suited as a bottom-up visual encoder for captioning. With the use of positional encoding, which we apply to the decoded tokens in the caption text, it can properly model sequential data. Rather of encoding an order for three things, our Object Relation Transformer aims to encode and weight how two objects are spatially connected to one another.

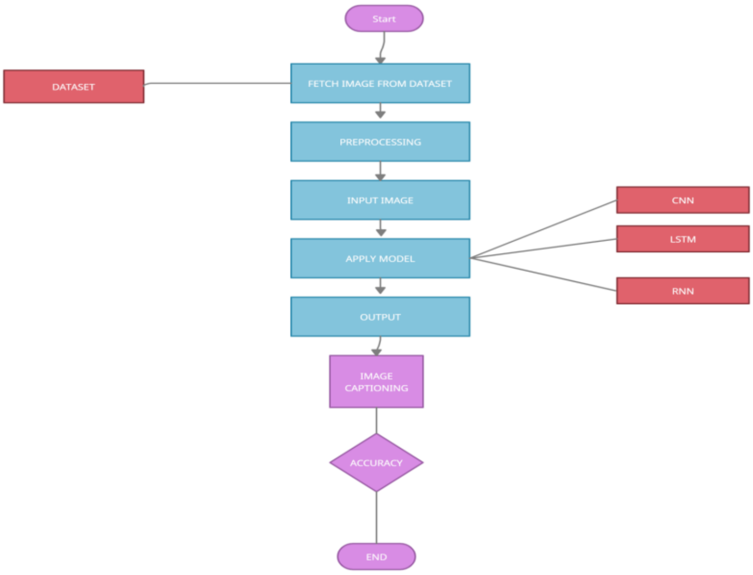
III. Methodology

This project will be developed using python, ML and web technology.

Expected Dataset: flicker 30k dataset.

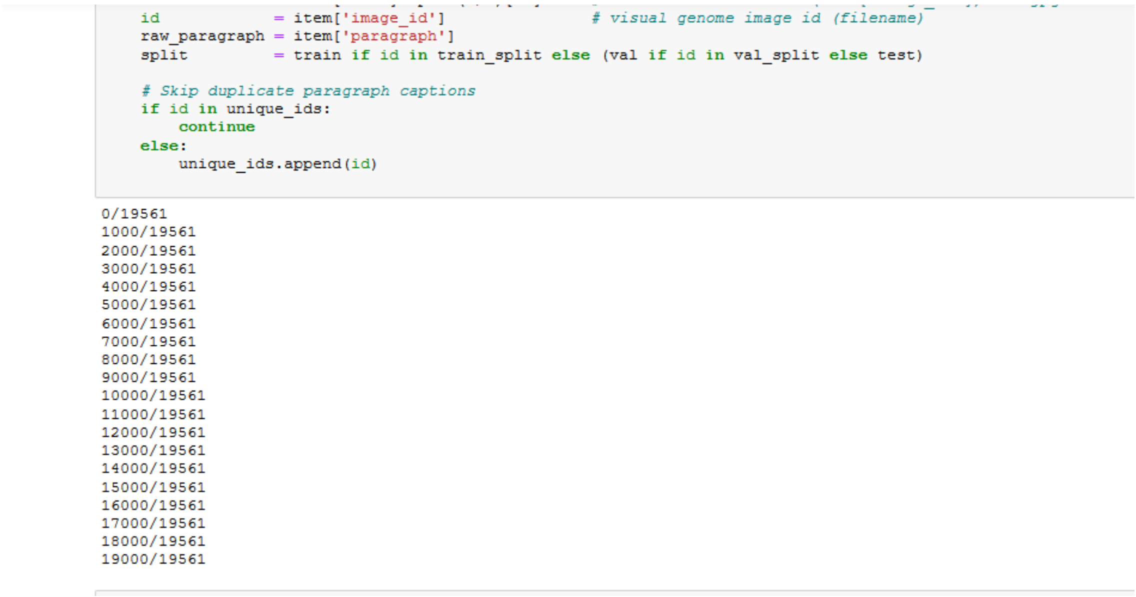
1. Image feature Detection: for these we are using a pre trained model based on transfer learning that is VGG16. It is already been installed in Keras library.
2. Text Generation using LSTM: long short term memory networks usually called LSTM's are a special kind of RNN. Capable of learning long-term dependencies
3. These all-purpose we are using python as backend, for training the model we are using ML algorithms.
4. We will implement an web application

The flow chart depicting the flow of our system is-



The dataset we used to train the model is as follows:

To begin the pre-processing, all uppercase terms were converted to lowercase. The following step was to eliminate all punctuation marks and emojis, as they were unnecessary for our purposes. The next step is to eliminate the stop words. "A," "on," and "all" are examples of popular words that aren't effective in detection. These terms have little meaning and are frequently omitted from manuscripts. A text file with offensive terms was prepared. If the tweet contains the harmful words, it is labeled as "TRUE," else it is labeled as "FALSE." The image and categorization are now stored in this final dataset, which will be used to implement the method..



```
    id            = item['image_id']              # visual genome image id (filename)
    raw_paragraph = item['paragraph']
    split         = train if id in train_split else (val if id in val_split else test)

    # Skip duplicate paragraph captions
    if id in unique_ids:
        continue
    else:
        unique_ids.append(id)
```

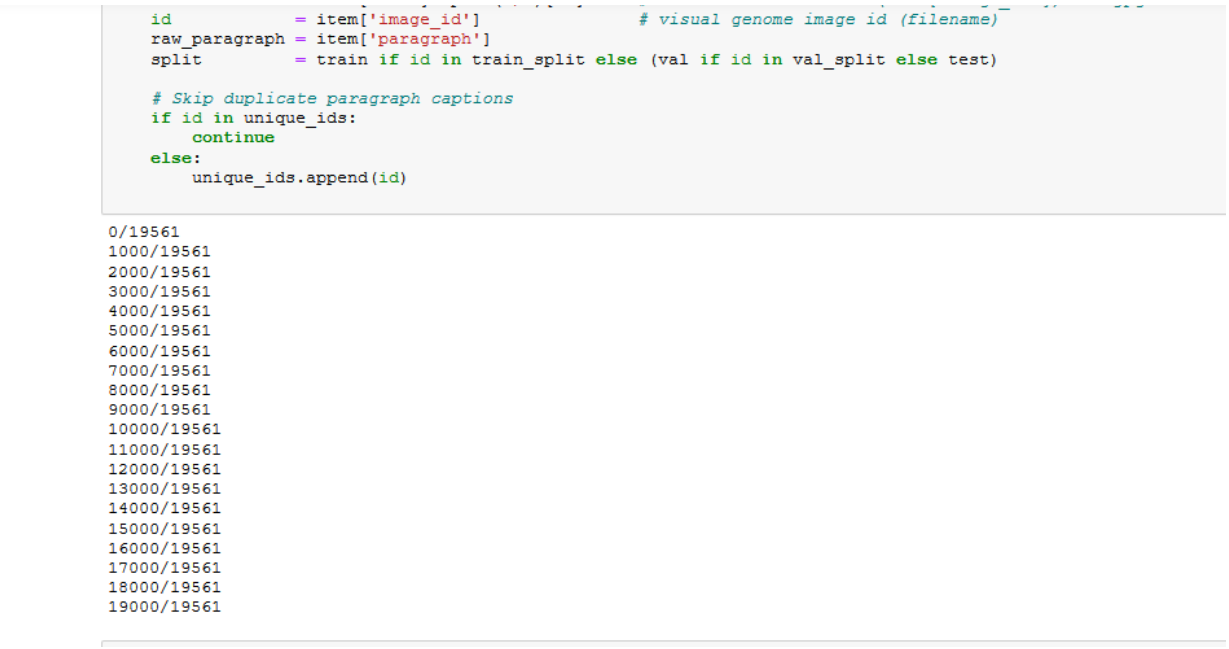
```
0/19561
1000/19561
2000/19561
3000/19561
4000/19561
5000/19561
6000/19561
7000/19561
8000/19561
9000/19561
10000/19561
11000/19561
12000/19561
13000/19561
14000/19561
15000/19561
16000/19561
17000/19561
18000/19561
19000/19561
```

Our proposed framework provides a feasible for Image Captioning. After completion of the project we will get a system which would successfully get the required text. The system will help the society especially the people who can just read and cannot understand images, It will also be feasible for Drivers that they can understand what is behind them to reduce chances of accident..

IV. Result Discussions

The system is successfully able to generatethe image captioning.

Prepro_captions.pyoutput

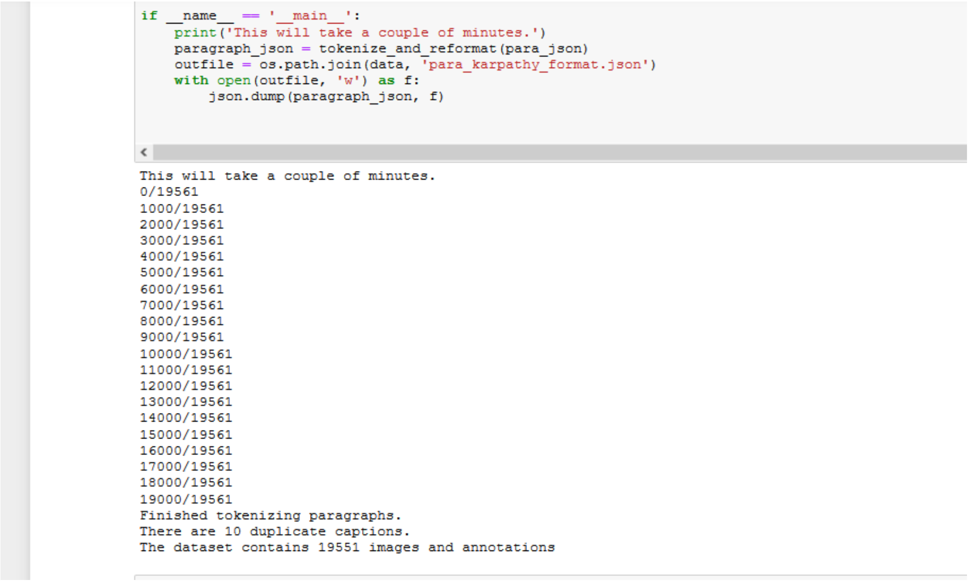


```
    id             = item['image_id']              # visual genome image id (filename)
    raw_paragraph = item['paragraph']
    split          = train if id in train_split else (val if id in val_split else test)

    # Skip duplicate paragraph captions
    if id in unique_ids:
        continue
    else:
        unique_ids.append(id)
```

```
0/19561
1000/19561
2000/19561
3000/19561
4000/19561
5000/19561
6000/19561
7000/19561
8000/19561
9000/19561
10000/19561
11000/19561
12000/19561
13000/19561
14000/19561
15000/19561
16000/19561
17000/19561
18000/19561
19000/19561
```

Fig4.6.1: Processingimages

Prepro_text.pyoutput

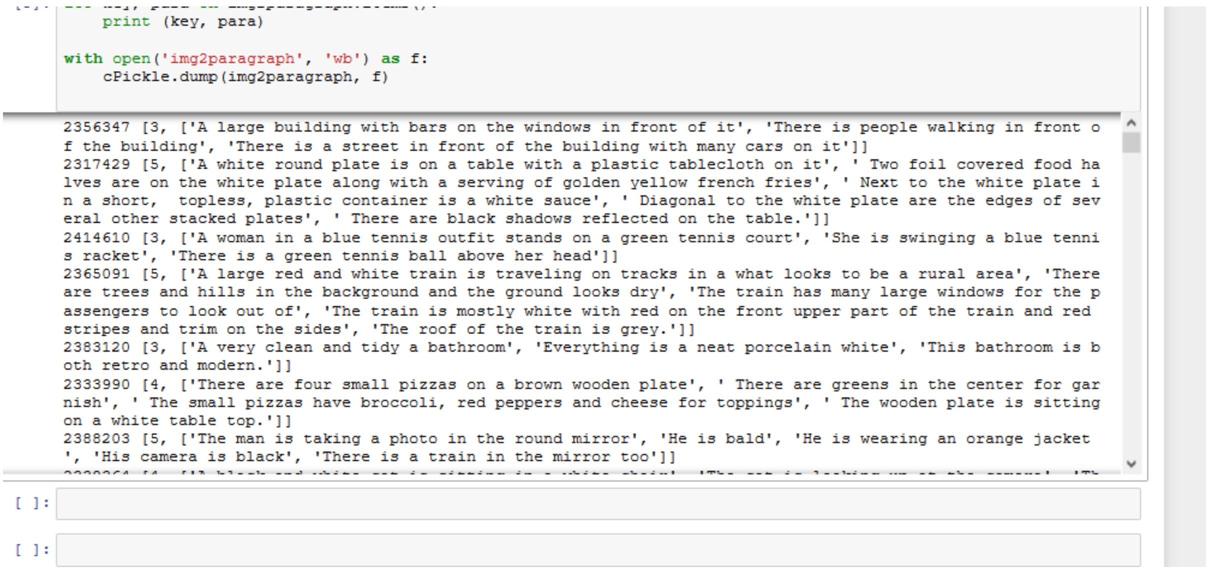


```
if __name__ == '__main__':
    print('This will take a couple of minutes.')
    paragraph_json = tokenize_and_reformat(para_json)
    outfile = os.path.join(data, 'para_karpathy_format.json')
    with open(outfile, 'w') as f:
        json.dump(paragraph_json, f)
```

```
This will take a couple of minutes.
0/19561
1000/19561
2000/19561
3000/19561
4000/19561
5000/19561
6000/19561
7000/19561
8000/19561
9000/19561
10000/19561
11000/19561
12000/19561
13000/19561
14000/19561
15000/19561
16000/19561
17000/19561
18000/19561
19000/19561
Finished tokenizing paragraphs.
There are 10 duplicate captions.
The dataset contains 19551 images and annotations
```

Fig4.6.2: Clearing Duplicates

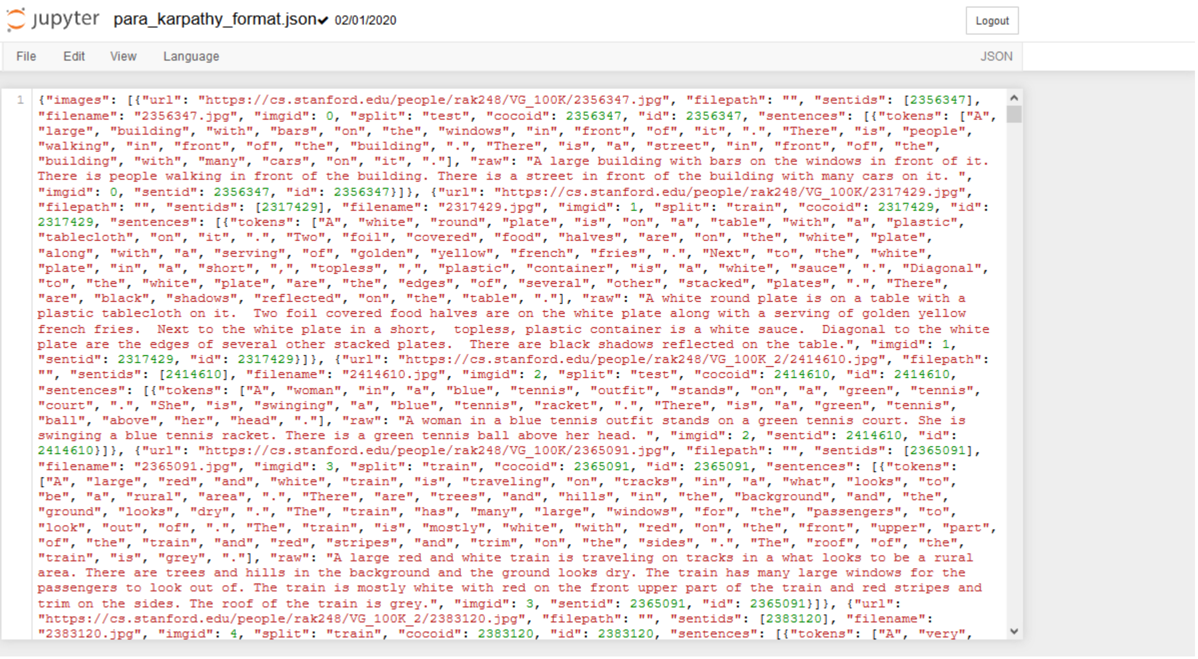
Parse-jsonoutput



```
    print (key, para)

with open('img2paragraph', 'wb') as f:
    cPickle.dump(img2paragraph, f)
```

```
2356347 [3, ['A large building with bars on the windows in front of it', 'There is people walking in front o
f the building', 'There is a street in front of the building with many cars on it']]
2317429 [5, ['A white round plate is on a table with a plastic tablecloth on it', ' Two foil covered food ha
lves are on the white plate along with a serving of golden yellow french fries', ' Next to the white plate i
n a short,  topless, plastic container is a white sauce', ' Diagonal to the white plate are the edges of sev
eral other stacked plates', ' There are black shadows reflected on the table.']]
2414610 [3, ['A woman in a blue tennis outfit stands on a green tennis court', 'She is swinging a blue tenni
s racket', 'There is a green tennis ball above her head']]
2365091 [5, ['A large red and white train is traveling on tracks in a what looks to be a rural area', 'There
are trees and hills in the background and the ground looks dry', 'The train has many large windows for the p
assengers to look out of', 'The train is mostly white with red on the front upper part of the train and red
stripes and trim on the sides', 'The roof of the train is grey.']]
2383120 [3, ['A very clean and tidy a bathroom', 'Everything is a neat porcelain white', 'This bathroom is b
oth retro and modern.']]
2333990 [4, ['There are four small pizzas on a brown wooden plate', ' There are greens in the center for gar
nish', ' The small pizzas have broccoli, red peppers and cheese for toppings', ' The wooden plate is sitting
on a white table top.']]
2388203 [5, ['The man is taking a photo in the round mirror', 'He is bald', 'He is wearing an orange jacket
', 'His camera is black', 'There is a train in the mirror too']]
```

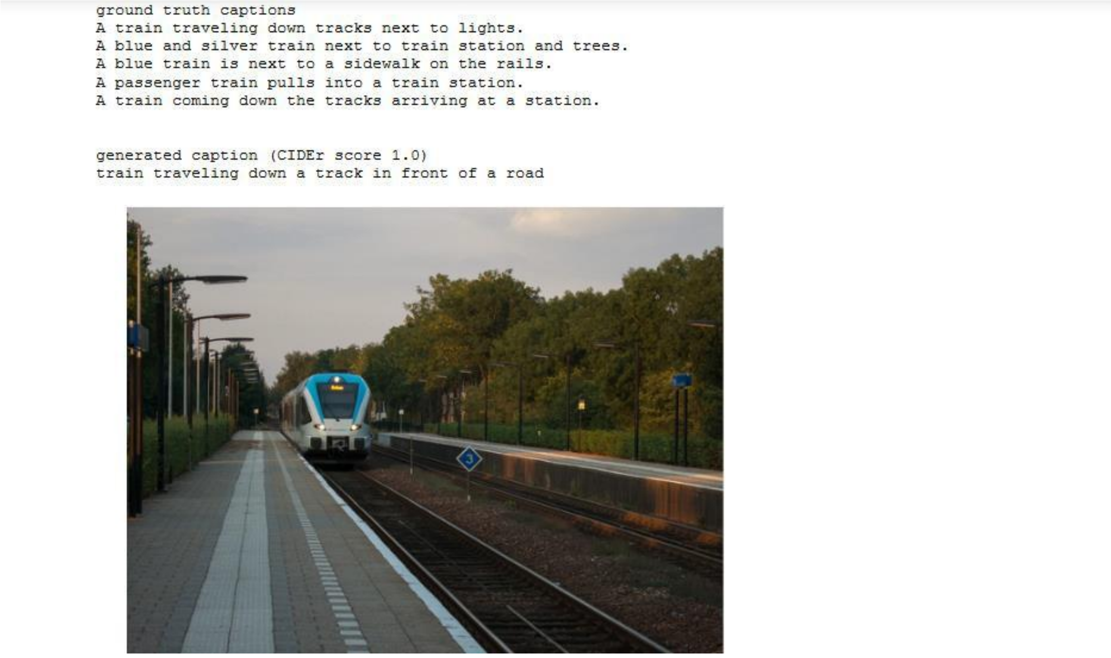
Fig4.6.3: Getting Paragraph



```
{"images": [{"url": "https://cs.stanford.edu/people/rak248/VG_100K/2356347.jpg", "filepath": "", "sentids": [2356347],
"filename": "2356347.jpg", "imgid": 0, "split": "test", "cocoid": 2356347, "id": 2356347, "sentences": [{"tokens": ["A",
"large", "building", "with", "bars", "on", "the", "windows", "in", "front", "of", "it", ".", "There", "is", "people",
"walking", "in", "front", "of", "the", "building", ".", "There", "is", "a", "street", "in", "front", "of", "the",
"building", "with", "many", "cars", "on", "it", "."], "raw": "A large building with bars on the windows in front of it.
There is people walking in front of the building. There is a street in front of the building with many cars on it. ",
"imgid": 0, "sentid": 2356347, "id": 2356347}]}, {"url": "https://cs.stanford.edu/people/rak248/VG_100K/2317429.jpg",
"filepath": "", "sentids": [2317429], "filename": "2317429.jpg", "imgid": 1, "split": "train", "cocoid": 2317429, "id":
2317429, "sentences": [{"tokens": ["A", "white", "round", "plate", "is", "on", "a", "table", "with", "a", "plastic",
"tablecloth", "on", "it", ".", "Two", "foil", "covered", "food", "halves", "are", "on", "the", "white", "plate",
"along", "with", "a", "serving", "of", "golden", "yellow", "french", "fries", ".", "Next", "to", "the", "white",
"plate", "in", "a", "short", ",", "topless", ",", "plastic", "container", "is", "a", "white", "sauce", ".", "Diagonal",
"to", "the", "white", "plate", "are", "the", "edges", "of", "several", "other", "stacked", "plates", ".", "There",
"are", "black", "shadows", "reflected", "on", "the", "table", "."], "raw": "A white round plate is on a table with a
plastic tablecloth on it.  Two foil covered food halves are on the white plate along with a serving of golden yellow
french fries.  Next to the white plate in a short,  topless, plastic container is a white sauce.  Diagonal to the white
plate are the edges of several other stacked plates.  There are black shadows reflected on the table.", "imgid": 1,
"sentid": 2317429, "id": 2317429}]}, {"url": "https://cs.stanford.edu/people/rak248/VG_100K_2/2414610.jpg", "filepath":
"", "sentids": [2414610], "filename": "2414610.jpg", "imgid": 2, "split": "test", "cocoid": 2414610, "id": 2414610,
"sentences": [{"tokens": ["A", "woman", "in", "a", "blue", "tennis", "outfit", "stands", "on", "a", "green", "tennis",
"court", ".", "She", "is", "swinging", "a", "blue", "tennis", "racket", ".", "There", "is", "a", "green", "tennis",
"ball", "above", "her", "head", "."], "raw": "A woman in a blue tennis outfit stands on a green tennis court. She is
swinging a blue tennis racket. There is a green tennis ball above her head. ", "imgid": 2, "sentid": 2414610, "id":
2414610}]}, {"url": "https://cs.stanford.edu/people/rak248/VG_100K/2365091.jpg", "filepath": "", "sentids": [2365091],
"filename": "2365091.jpg", "imgid": 3, "split": "train", "cocoid": 2365091, "id": 2365091, "sentences": [{"tokens":
["A", "large", "red", "and", "white", "train", "is", "traveling", "on", "tracks", "in", "a", "what", "looks", "to",
"be", "a", "rural", "area", ".", "There", "are", "trees", "and", "hills", "in", "the", "background", "and", "the",
"ground", "looks", "dry", ".", "The", "train", "has", "many", "large", "windows", "for", "the", "passengers", "to",
"look", "out", "of", ".", "The", "train", "is", "mostly", "white", "with", "red", "on", "the", "front", "upper", "part",
"of", "the", "train", "and", "red", "stripes", "and", "trim", "on", "the", "sides", ".", "The", "roof", "of", "the",
"train", "is", "grey", "."], "raw": "A large red and white train is traveling on tracks in a what looks to be a rural
area. There are trees and hills in the background and the ground looks dry. The train has many large windows for the
passengers to look out of. The train is mostly white with red on the front upper part of the train and red stripes and
trim on the sides. The roof of the train is grey.", "imgid": 3, "sentid": 2365091, "id": 2365091}]}, {"url":
"https://cs.stanford.edu/people/rak248/VG_100K_2/2383120.jpg", "filepath": "", "sentids": [2383120], "filename":
"2383120.jpg", "imgid": 4, "split": "train", "cocoid": 2383120, "id": 2383120, "sentences": [{"tokens": ["A", "very",
```

Fig4.6.4: Tokenization

OUTPUT



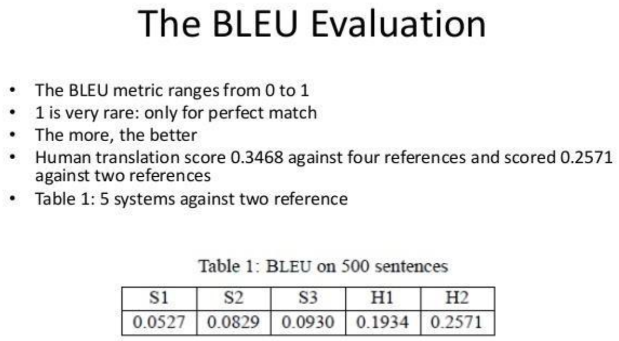
```
ground truth captions
A train traveling down tracks next to lights.
A blue and silver train next to train station and trees.
A blue train is next to a sidewalk on the rails.
A passenger train pulls into a train station.
A train coming down the tracks arriving at a station.


generated caption (CIDEr score 1.0)
train traveling down a track in front of a road
```

Fig Generating Captions

ANALYSIS

BLEU -

BLEU is a quality metric score for MT systems that attempts to measure the correspondence between a machine translation output and a human translation. The central idea behind BLEU is that the closer a machine translation is to a professional human translation, the better it is.



The BLEU Evaluation

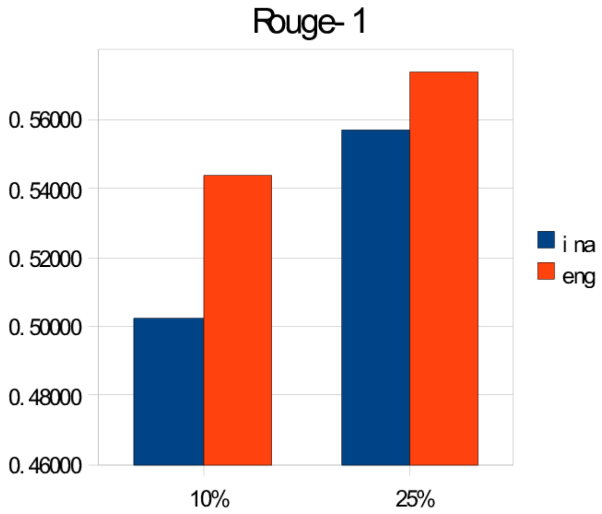
- The BLEU metric ranges from 0 to 1
- 1 is very rare: only for perfect match
- The more, the better
- Human translation score 0.3468 against four references and scored 0.2571 against two references
- Table 1: 5 systems against two reference

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

ROUGE -

ROUGE(Lin, 2004) is initially proposed for evaluation of summarization systems, and this evaluation is done via comparing overlapping n-grams, word sequences and word pairs. In this study, weuse ROUGE-Lversion, which basically measures the longest common subsequences between a pairof sentences.SinceROUGEmetric relies highly on recall, it favors long sentences, as also notedby (Vedantam et al., 2015).

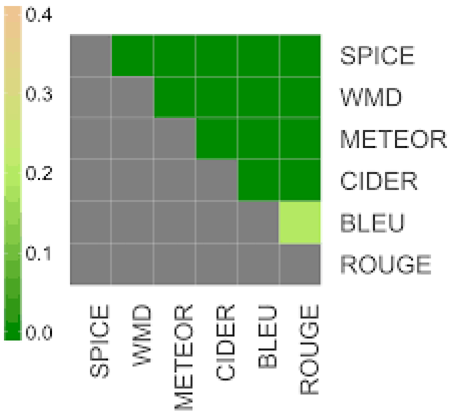


Graph on Rouge

CIDEr-

CIDEr(Consensus-based Image Description Evaluation). This metric measures the similarity of a generated sentence against a set of ground truth sentences written by humans. This metric shows high agreement with consensus as assessed by humans.

.



Graph comparing metrics

ACCURACY

We employ the human consensus scores while evaluating the accuracies .In particular, for evaluation, a triplet of descriptions, one reference and two candidate descriptions, is shown to human subjects and they are asked to determine the candidate description that is more similar to the reference. A metric is ac-curate if it provides a higher score to the description chosen by the human subject as being more similar to the reference caption

TESTING

Figure:Testing Image1

Paragraph: There are five horses running on the road.The horses are running one by one in a line. The sky is blue and cloudy. The horses are white and brown in colour. A poles are present at a far distance from horses.

.

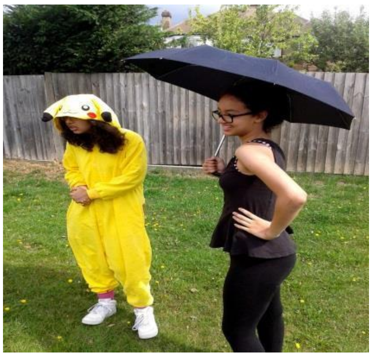


Figure 4 :Testing Image 2

Paragraph:There is a girl and a person stand on the grass.A person wear a black shirt and black pant keep black glasses hold umbrella.A yellow dress girl have a white shoes with acap.A long trees are there behind the two person

SENTENCES	BLEU	CIDER	ROUGE	METEOR
S1	0.579	0.600	0.396	0.195
S2	0.404	0.658	0.274	0.256
S3	0.279	0.599	0.400	0.172
S4	0.191	0.677	0.450	0.137

.

Table :Experiment Result: The sentences generated are evaluated under differentmetrics with respect to a sentreference ence by calculating similarity of every word.

Metrics	LSTMModel	GRU	RNN
BLEU	0.579	0.473	0.525
CIDER	0.600	0.598	0.600
ROUGE	0.396	0.333	0.354
METEOR	0.195	0.188	0.190

Different methods are performed against different evaluation metrics.LSTM is better than other methods[6]

V. Conclusion

As can be seen from the study, finding a segmentation method that adapts to allimages is tough. At the moment, image segmentation theory research isn't ideal, and there are still a lot of practical issues in applied research. The following trends in image segmentation techniques may emerge from a comparison of the benefits and drawbacks of several picture segmentation algorithms:

1) The use of a variety of segmentation methods. Because of the image's diversity and unpredictability, it's vital to combine numerous segmentation approaches and fully use the benefits of various algorithms based on multi-feature fusion in order to produce better segmentation results.

2) We will learn about the VGG16 model architecture, the Long Short Term Memory Network, how to merge these two models, and the blue score in this project. What is the process by which the LSTM network creates captions? How can we apply the VGG16 model for our project, and how can we leverage deep learning to generate captions from images?

Acknowledgment

We as the authors would like to extend a special thanks of vote to the reviewers of this paper for their valuable suggestions to improve this paper.
The paper is supported by Acropolis Institute of Technology and Research,Indore (M.P.)

REFERENCES

[1] Marc'AurelioRanzato, Sumit Chopra, Michael Auli,andWojciechZaremba. 2015.‖ Sequence level training with recurrent neural networks‖arXiv preprintarXiv:1511.06732.

[2] Peter Anderson, Xiaodong He, Chris Buehler, DamienTeney, Mark Johnson, Stephen Gould, and LeiZhang. 2017.‖ Bottom-up and top-down attention for image captioning and vqa‖. arXiv preprintarXiv:1707.07998.

[3] Steven J Rennie, Etienne Marcheret, Youssef Mroueh,Jarret Ross, and VaibhavaGoel. 2016‖. Self-critical sequence training for image captioning.‖

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, JamesHays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014.‖ Microsoft coco:Common objects in context.‖ In European conference on computer vision, pages 740–755. Springer.

[5] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan,and Eric P Xing. 2017. Recurrent topic-transition for visual paragraph generation.‖

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: ―Generating sentences from images.‖ InECCV,2010.

[7] Karpathy and L. Fei-Fei.‖ Deep visual-semantic generating image descriptions.‖InCVPR,2015.

[8] G.Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg,and T. L. Berg. Baby talk: ―Understanding and generating image descriptions.‖InCVPR, 2011.

[9] J. Donahue,L. Anne Hendricks,S. Guadarrama,M. Rohrbach, S. Venugopalan, K. Saenko, and T.Darrell‖Long-term recurrent convolutional networks for and description‖.

[10] InCVPR, 2015.Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis,and Erkut Erdem. 2016. Re-evaluating automaticmetrics for image captioning.arXiv preprintarXiv:1612.07600.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, andDumitru Erhan. 2014. Show and tell: A neural im-age caption generator.CoRR, abs/1411.4555.

[12] Ruotian Luo, Brian L. Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. CoRR,abs/1803.04376.

[13] Steven J Rennie,Etienne Marcheret,Youssef Mroueh,Jarret Ross,and Vaibhava Goel.2016"Self Critical Sequence Training for Image Captioning".

[14] Tsung-Yi Lin,Michael Maire,Serge Belongie,James Hays,Pietro Peronal Deva Ramanan,Piotr Dollar, and C Larence Zitnick.2014, "Micrososococo:Common Objects in ConteAt"In European conference on computer vision pages 740-755.Springer.

[15] Xiaodan Liang,Zhiting Hu,Hao Zhang,Chuang Ggan ,and Eric P Xing.2017. "Recurrent Topic-Transition for Visual Paragraph Generation."