Image Captioning using Attention Mechanism with ResNet, VGG and Inception Models

Ms. Shruti Mundargi[1], Mr. Hrushikesh Mohanty[2]

[1]Student, Department of Computer Engineering, Excelsior Education Society's, K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India
[2]Student, Department of Information Technology, Excelsior Education Society's, K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India

---***---

Abstract - *With increasing advancements in the field of artificial intelligence, image captioning has become a popular topic because of its wide range of uses and the fact that it is the result of the union of two most interesting and important topics of artificial intelligence, computer vision and natural language processing. In this paper, we have conducted an analysis using visual attention mechanism in image captioning to find out the efficiency and accuracy of the output captions using three preprocessing models InceptionV3, VGG19 and ResNet50. Along with this, we have also given a brief about the concepts that are relevant to image captioning and used in our experiment. In order to perform the analysis we have used images from MS-COCO dataset.*

***Key Words*: LSTM, GRU visual attention mechanism, CNN encoder-decoder, image captioning.**

1. INTRODUCTION

Image captioning is a process in which the input image is determined and a description is provided in the form of caption using natural language processing. Image captioning is not only the application of computer vision but also the application of the integration of computer vision and natural language processing. It is the primary implementation of contextual understanding of an image and text generation based on that understanding. This is the reason why image caption generation has been a very popular challenge for machine learning algorithms. There have been many different advances in the field of computer vision and natural language processing which has improved the accuracy of image caption generating models in general. This process proves to be a very important technology and has several applications. As self-driving cars are becoming the next big thing, object detection and forming relations between the detected objects used to understand the scenario would be crucial. Image captioning can not only cost the entire self-driving system but also an integration of image captioning and text to speech would result in building communication between the machine and passengers. Also, the CCTV cameras around the world could use image captioning to detect and convey alarming situations, help in investigating crimes efficiently and act as 'smart guards'. An AI could be built to help the visually impaired using image captioning. It would also help search engines come up with better image search results based on the input given by the user.

There are several image captioning methods like template-based image captioning, retrieval-based image captioning Template-based methods have predefined templates with some blank slots to generate captions. In these approaches, different objects, attributes, actions are detected first and then the blank spaces in the templates are filled. However, templates are predefined and cannot generate variable-length captions. Moreover, later on, parsing based language models have been introduced in image captioning which are more powerful than fixed template-based methods. In retrieval-based approaches, captions are retrieved from a set of existing captions [3]. Retrieval based methods first find the visually similar images with their captions from the training data set. These captions are called candidate captions. The captions for the query image are selected from these captions pool. These methods produce general and syntactically correct captions. However, they cannot generate image specific and semantically correct captions [3].

2. RELATED WORK

There have been extensive studies done related to computer vision and image recognition and captioning. The significant work done in visual attention was first proposed in the paper, '*Attend and Tell: Neural Image Caption Generation with Visual Attention*' [1]. The authors explain how visual attention would prove to be much more effective and efficient for image captioning problems. Their model has been trained using a deterministic manner and standard back propagation techniques. The concept of visual attention with encoder-decoder approach is described in this work. This is the similar approach that we have followed in our study.

Along with that, our study was heavily influenced by the paper '*Multimodal Neural Language Models*' [2]. The authors here have introduced multimodal neural language models so that models of natural language processing can be conditioned to other modalities. These models were experimented on two datasets - 'IAPR TC-12' and 'Attribute

Discovery', each containing 20,000 images and 40,000 images, respectively.

3. IMAGE CAPTIONING USING DEEP LEARNING MODELS

Deep Learning:

Deep Learning is a subset of machine learning and is related to algorithms for artificial neural networks inspired by the structure and working of a living brain. These neural networks consist of multiple layers which extract features from the dataset fed to them and analyse these features by giving them certain weights [3]. Traditionally, layers that appear later in the neural net model extract and analyse higher or more complex features. For example, in image recognition models, lower layers would either check the edges or borders and size of the images, whereas the higher layers would try to identify the subjects, object borders, categories of the objects present in the image or identify the letters and characters spotted inside the image.

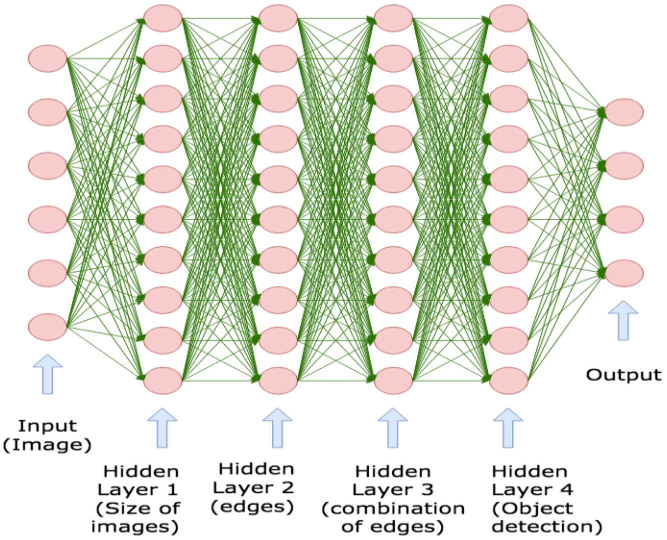

Input (Image)
Hidden Layer 1 (Size of images)
Hidden Layer 2 (edges)
Hidden Layer 3 (combination of edges)
Hidden Layer 4 (Object detection)
Output

Fig-1: Deep neural network for image recognition

Long Short Term Memory (LSTM):

LSTM is a type of Recurrent Neural Network or RNN in short. RNNs have a chain-like structure where the data or input of the previous layer plays a part in determining the output of the next layer of the model. These RNNs could be used in determining the words in the sentence by checking all the previous words that came before in the same sentence. LSTM stands for Long Short Term Memory and they are specifically designed to remember information for a long period of time by default. They are a special kind of RNN as they can retain both short-term and long-term memory. LSTMs have three main components in their cell, each for forgetting, remembering and updating data. They also have a chain-like structure like RNN with the difference being that in LSTM, instead of a single neural net layer, they have four layers interacting with each other uniquely [5].

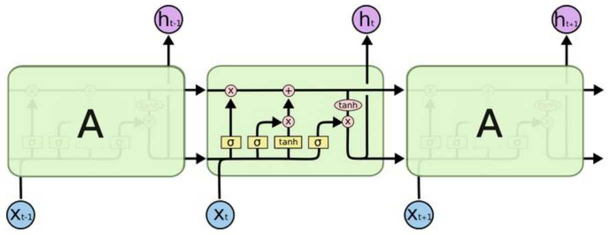


Fig -2: LSTM layer (https://colah.github.io/posts/2015-08-Understanding-LSTMs/)

BiLSTM can be said to be a version of LSTM itself. As stated before, in LSTM the previous layers inputs and outputs play a part in determining the output of the present layer. In BiLSTM, this is taken to a next level by checking the input of the following layer as well. For example, in the sentences, "The **train** arrived early at the station" and "Muhammad Ali asked his coach to **train** him in martial arts as well", the highlighted word "train" has different meanings. Therefore to predict a particular word, the context of the sentence is to be determined first and to predict the context, one has to look at the words that come before and after the said word. BiLSTM does just that by checking the previous word and the following word to check the context of the word and then predict the word by applying the forward propagation 2 times, one for forward cells and one for the backward cells [4].

Above we saw that LSTM uses something called Encoder and Decoder for input and output sequence. Using these LSTM transforms one sequence into a completely different sequence. To understand this, consider an example of an image captioning model. The encoder and decoder have their own imaginary language to communicate with each other, so this way, the encoder understands what is present in the input image as well as the imaginary language and the decoder understands the output and the imaginary language. So when the input is an image a CNN encoder generates a hidden state. This hidden state is decoded using an LSTM model which acts as our decoder, and a caption is generated word by word. But with all this, the context is also important. This happens just like a human sees a picture and understands what is happening in the picture. While looking at a picture a human would also keep in mind all the aspects of the image just to remember the context of the current word.

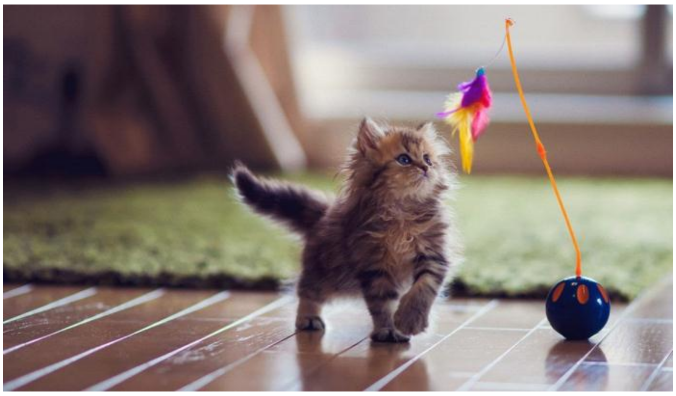


Fig-3: Example of encoder-decoder model works

For example, fig-3 would be described as "*Cat playing with a toy*" if a human were to describe it. This exact same process happens with Encoder and Decoder as well. While translating the image to the output description, the Decoder will also remember the important aspects and their arrangement in that image. LSTM does this by assigning priority or weights and checking the context every time it gives an output.

Gated Recurrent Unit (GRU):

GRUs are like LSTMs in many ways as both were introduced to get rid of the vanishing gradient problem.

In the below figure the *ht* holds information about the current unit and *ht-1* holds information for *t-1* units *zt* is the update gate and *rt* is the reset gate. *tanh* is the *tanh* activation function.

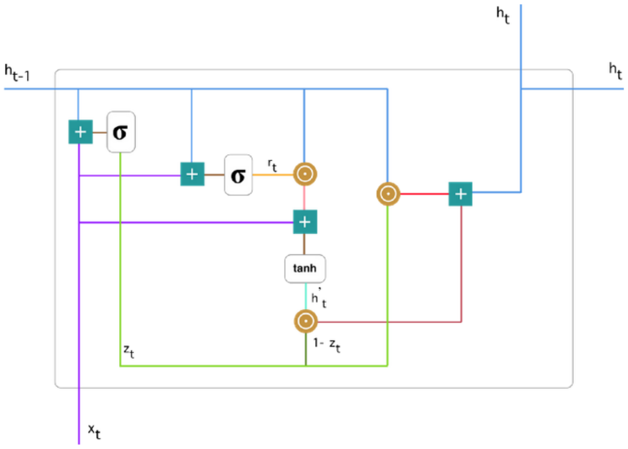


Fig-4: Gated Recurrent Unit (https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be)

GRU consists of an update gate and a reset gate. The update gate in GRUs decides how much the unit updates its activation or content. Basically, update gates decide how much data or information is to be passed for future reference [6]. Although in LSTM there is a mechanism which can control the degree of state to be exposed but GRUs expose the entire state each time. The formula for it is, where *zt* is the update gate: [7]

$$z_t^j = \sigma\left(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1}\right)^j$$

The reset gate is used to forget the past information in the model. It makes the unit act as if it is reading the first words of input sequence and makes it forget the earlier state. The formula for it is, where *rt* is the reset gate: [7]

$$r_t^j = \sigma\left(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1}\right)^j$$

Like LSTM GRUs gives excellent results and can be used with large number of layers

Visual Attention:

In the attention mechanism, the input image is divided into 'n' number of sections. So, when CNN encoders the hidden state it generates 'n' number of hidden states each representing its respective section. While decoder is generating words, it only focuses on the relevant part of the image.

This method works efficiently and avoids any irrelevant data to be included as description. There are two main types of attention mechanisms that are local (Bahdanau Attention) and global attention (Luong's Attention). In global attention all the 'n' number of sections are considered by the RNN decoder for generating words.

Whereas in local attention some of the sections are selected for the caption generation [9].

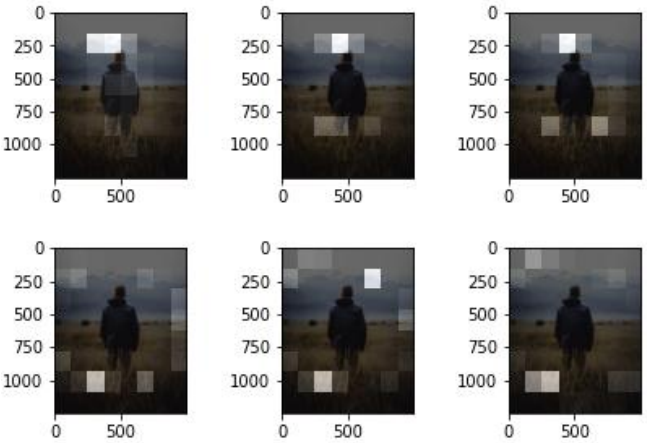


Fig-5: Example of local attention mechanism

Consider figure-5, only relevant sections have been considered and highlighted accordingly by the model in order to generate the captions.

How does the attention mechanism work?

Fig-6: Working of attention mechanism

Considering the above figure, in order to generate a caption for it, one would have to consider the important aspects of the image. Firstly, the section with the dog would be taken into consideration. Then the section around it, that is, the section containing purple flowers. Thus the caption generated here would be *"Dog between purple flowers."*

In our experiment we have used the local attention mechanism in order to generate the captions. The features are extracted from the lower convolutional layers from the pre-processing models InceptionV3, ResNe50 and VGG19, which gave us a vector shape of (8, 8, 2048). This is then reshaped to (64, 2048). This vector is then passed to a CNN single fully connected layer (encoder). We have used GRU as a RNN decoder for predicting the captions.

4. RESULTS AND ANALYSIS

In this study we have used the MS-COCO dataset. MS-COCO consists of 82,000 images which is used for several image classification tasks. We will be using 40,000 images from this dataset [8]. We performed training on 3 models, InceptionV3, ResNet50, and VGG19, to check which model gives better results.

About InceptionV3, ResNet50 and VGG19

InceptionV3 is an image recognition model which has accuracy more than 78.1% on ImageNet dataset. The model consists of several symmetric and asymmetric building blocks that include convolutions, dropouts, concats, pooling and fully connected layers. We extracted features from the last convolutional layer and created a model where the output layer is the last convolutional layer in the InceptionV3 architecture [10].

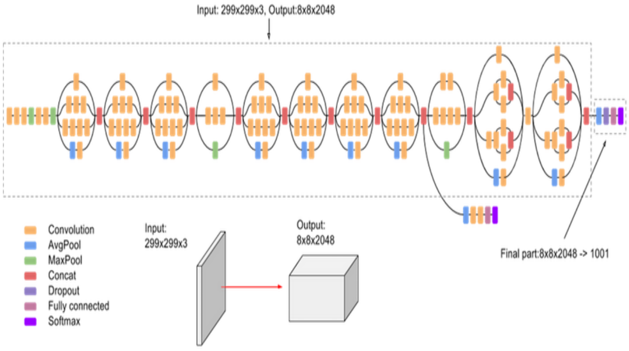


Fig-7: InceptionV3 Architecture (https://cloud.google.com/tpu/docs/inception-v3-advanced)

ResNet50 is a 50 layer deep convolutional neural network. It is trained on millions of images from the ImageNet dataset. ResNet is highly used for various computer vision tasks. It has the ability to classify images into 1000 categories.

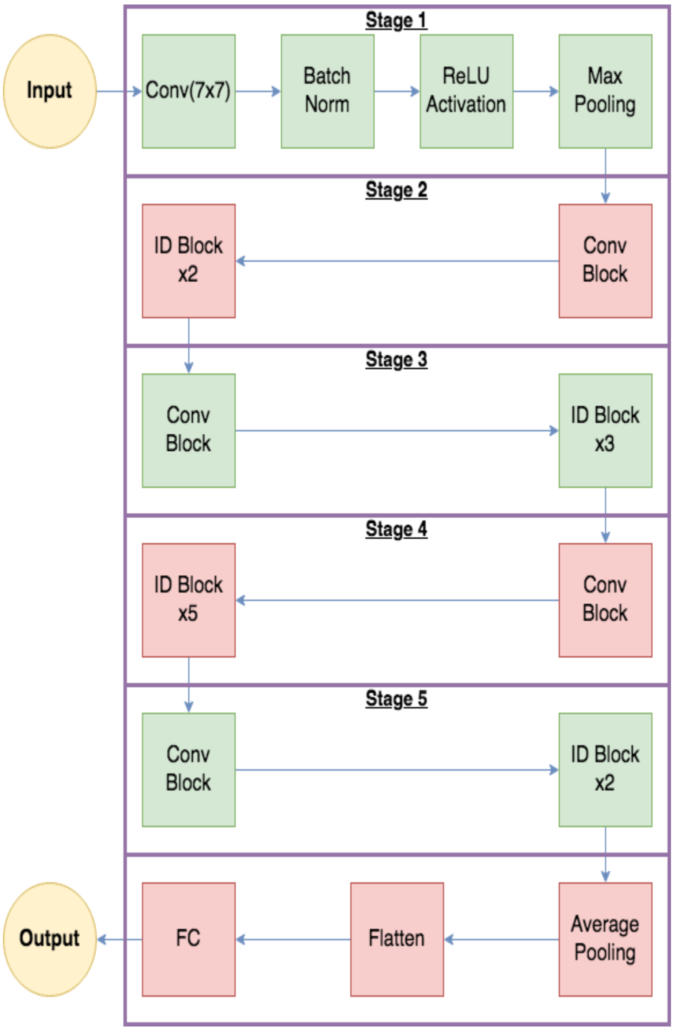


Fig-8: ResNet50 Model

VGG19 consists of 19 layers which has 16 convolutional layers, 5 maxpool layers, 3 fully connected layers and 1 softmax layer.

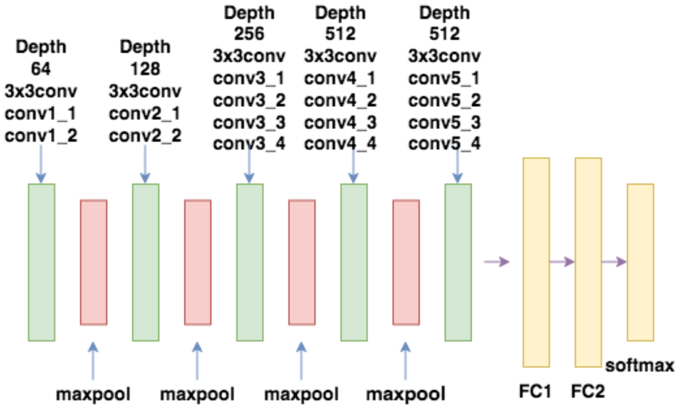


Fig-9: VGG19 Architecture

The results for the image captioning models were as following:

For InceptionV3

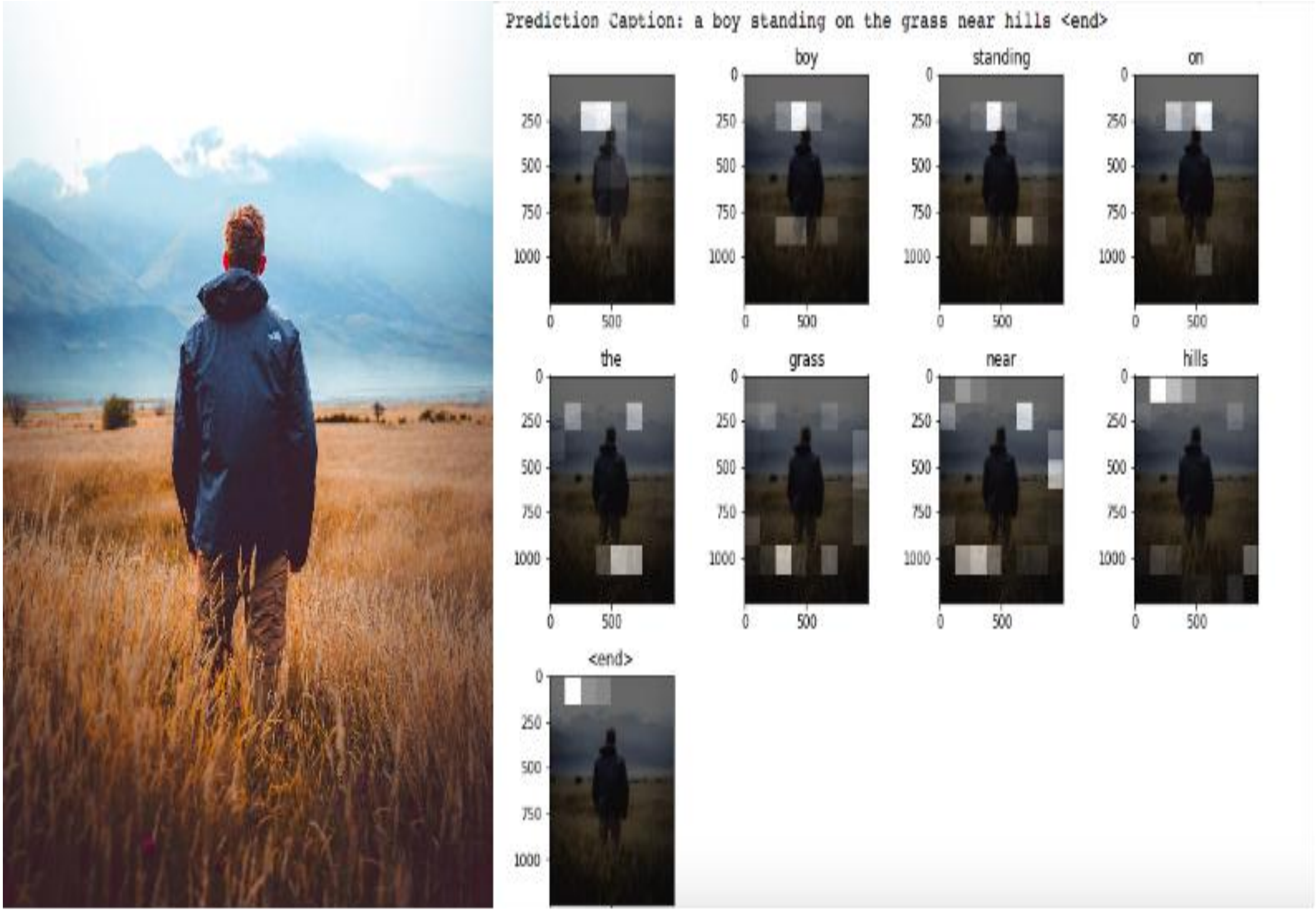

Prediction Caption: a boy standing on the grass near hills <end>
boy
standing
on
the
grass
near
hills
<end>
0
250
500
750
1000
0
500

Fig-10: Image 1 with predicted caption "*a boy standing on grass near hills*"

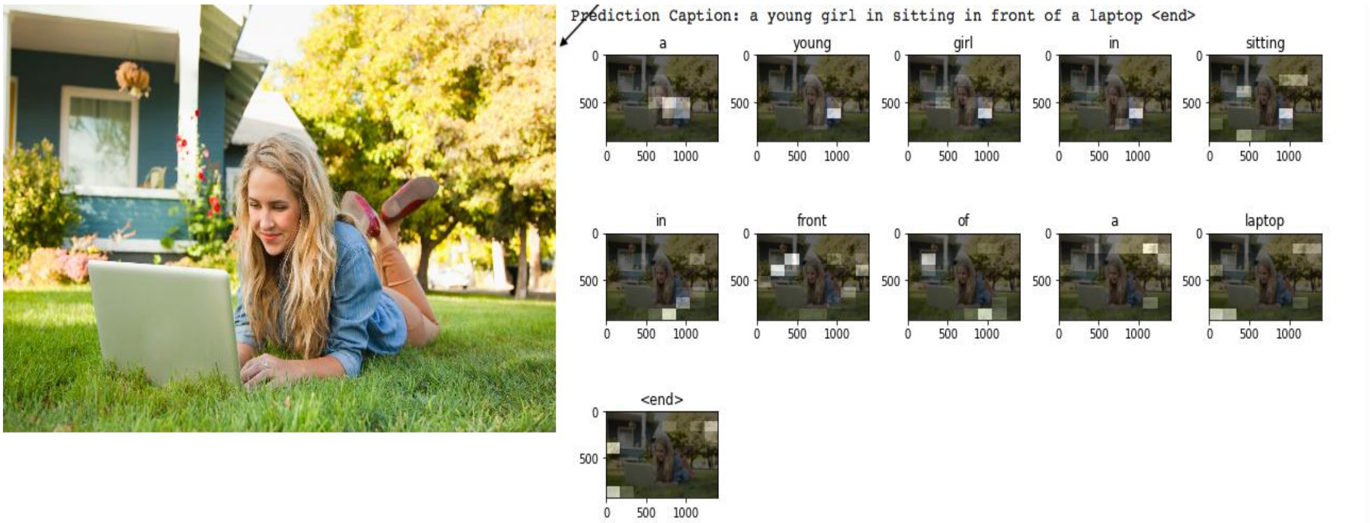

Prediction Caption: a young girl in sitting in front of a laptop <end>
a
young
girl
in
sitting
in
front
of
a
laptop
<end>
0
500
1000

Fig-11: Image 2 with predicted caption *"A young girl in sitting in front of a laptop"*

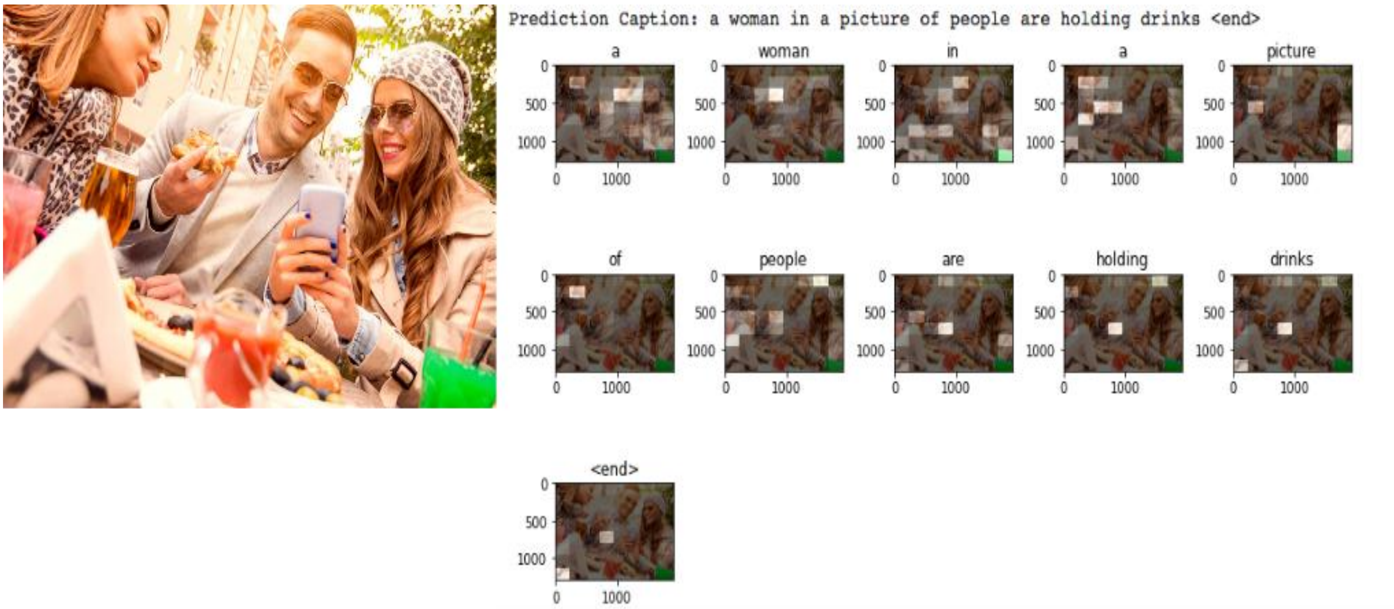

Prediction Caption: a woman in a picture of people are holding drinks <end>
a
woman
in
a
picture
of
people
are
holding
drinks
<end>
0
500
1000

Fig-12: Image 3 with predicted caption *"a woman in a picture of people are holding drinks"*

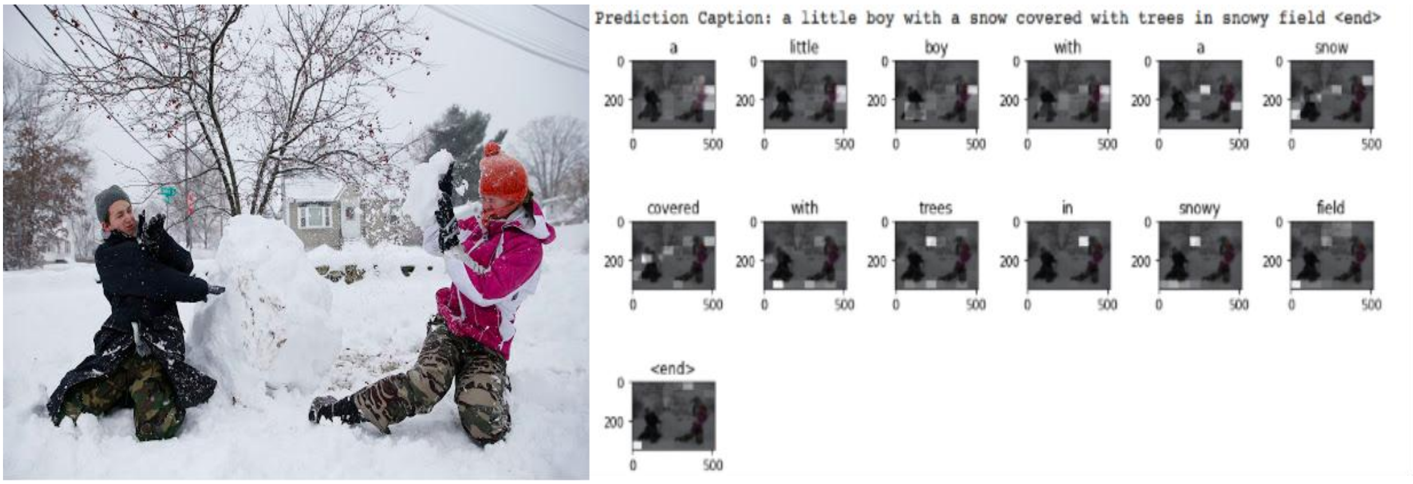
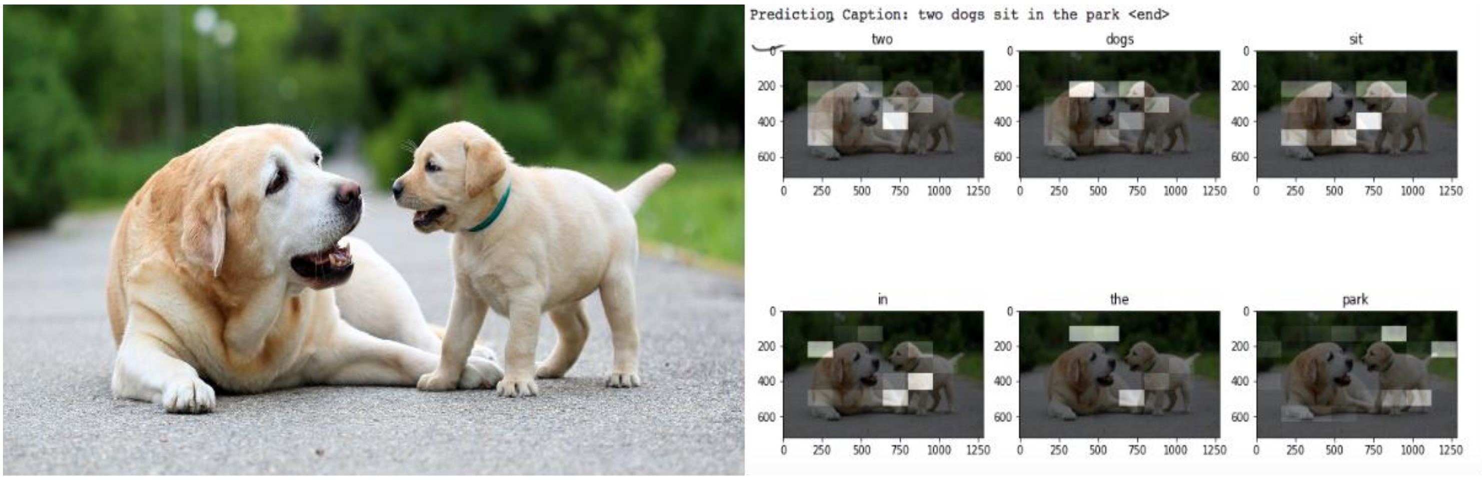

Prediction Caption: a little boy with a snow covered with trees in snowy field <end>
a
little
boy
with
a
snow
covered
with
trees
in
snowy
field
<end>
0
200
500

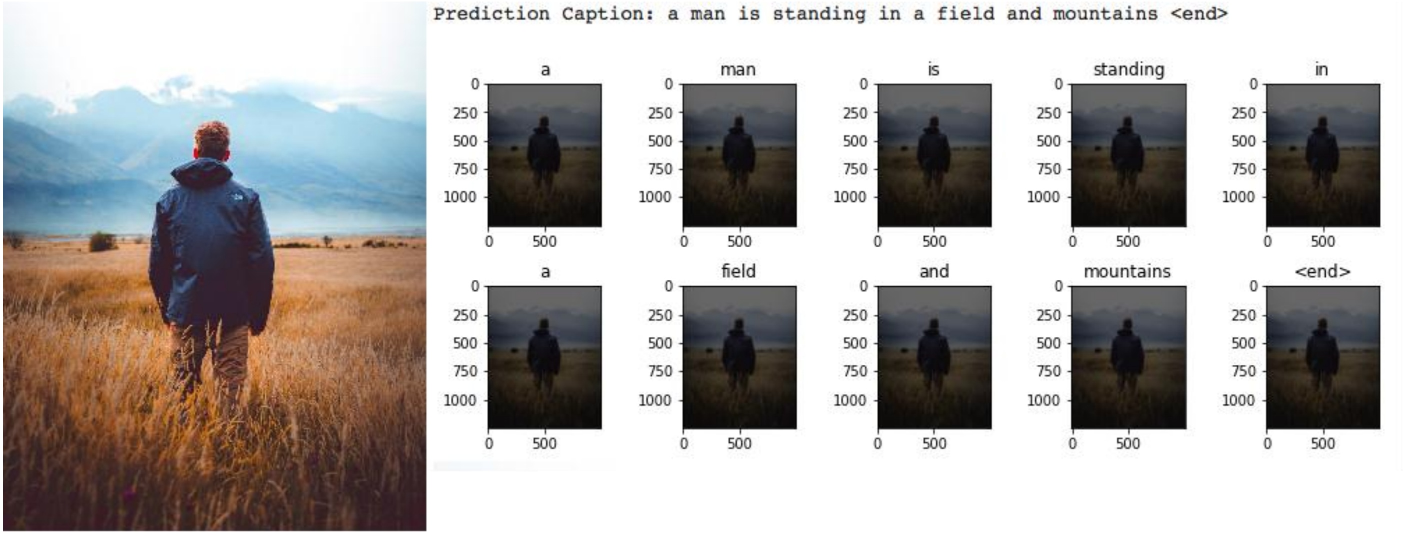
Fig-13: Image 4 with the predicted caption *"a little boy with snow covered with trees in snowy field."*



```
Prediction Caption: two dogs sit in the park <end>
```

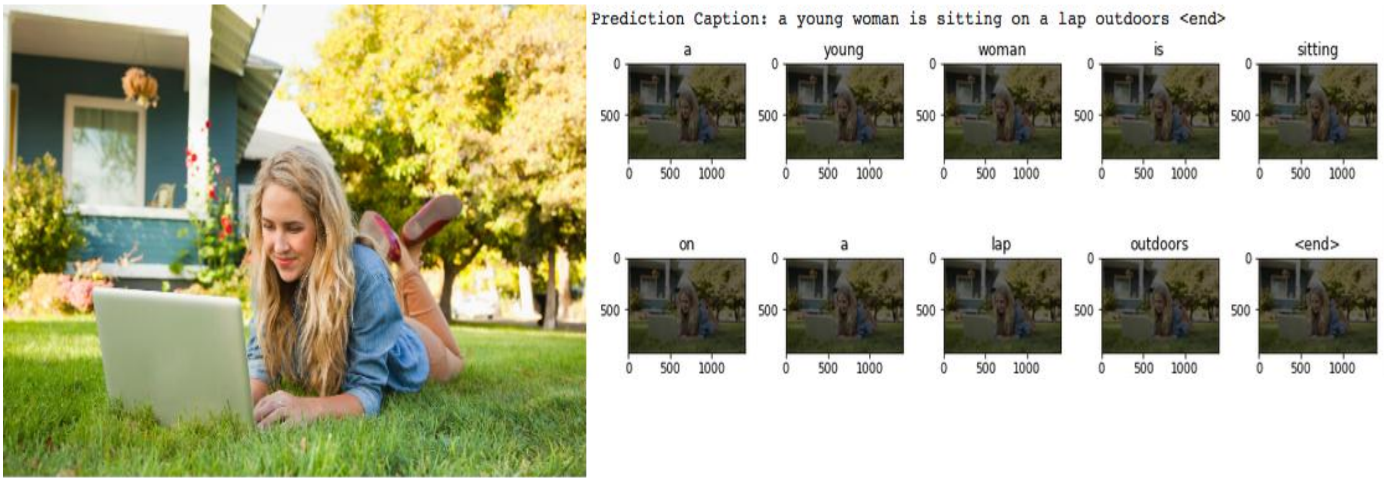
Fig-14: Image 5 with predicted caption *"two dogs sit in park"*

For ResNet50



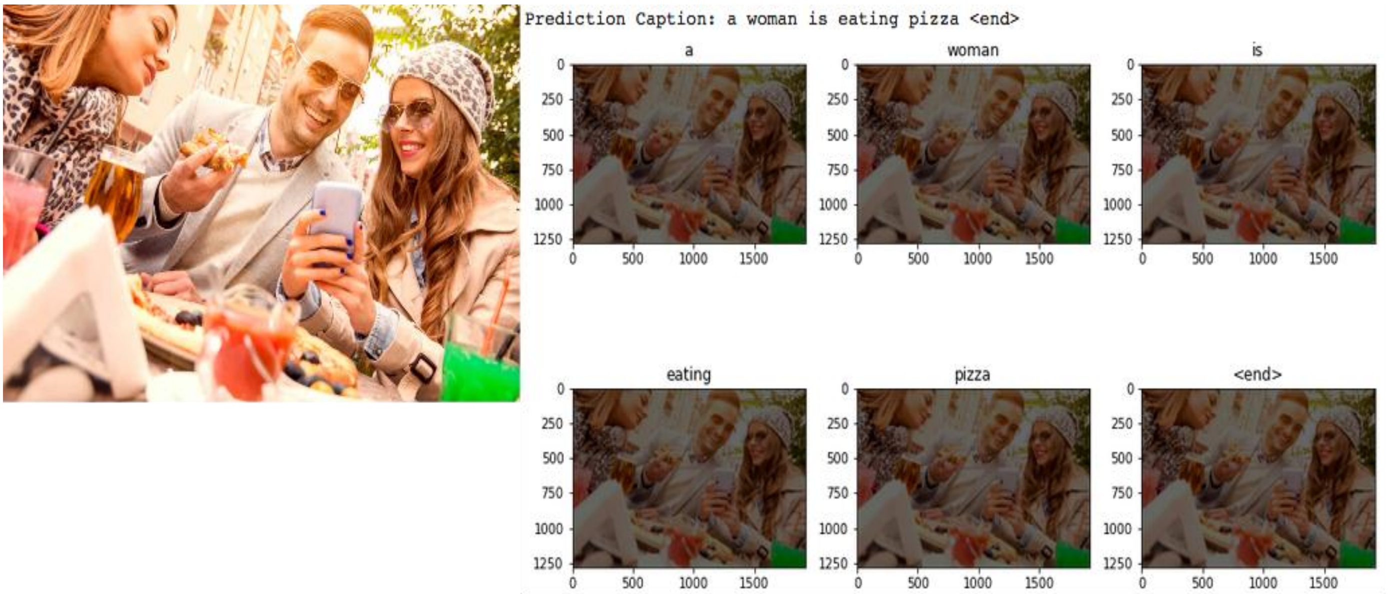
```
Prediction Caption: a man is standing in a field and mountains <end>
```

Fig-15: Image 1 with predicted caption *"a man standing in a field and mountains"*



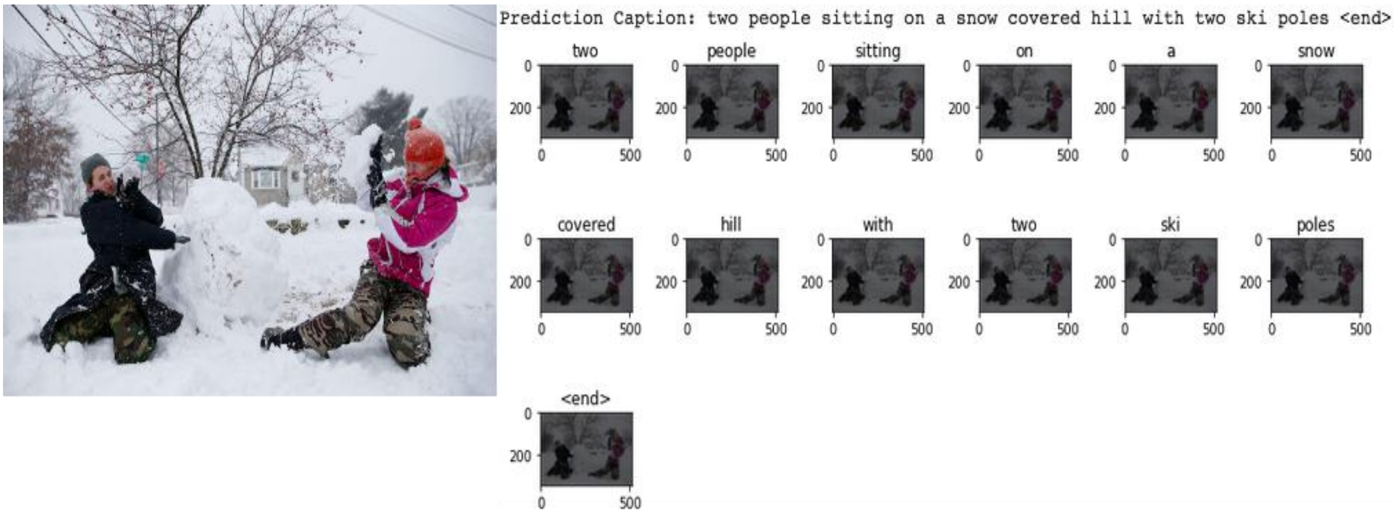
```
Prediction Caption: a young woman is sitting on a lap outdoors <end>
```

Fig-16: Image 2 with predicted caption *"a young woman is is sitting on a lap outdoors"*



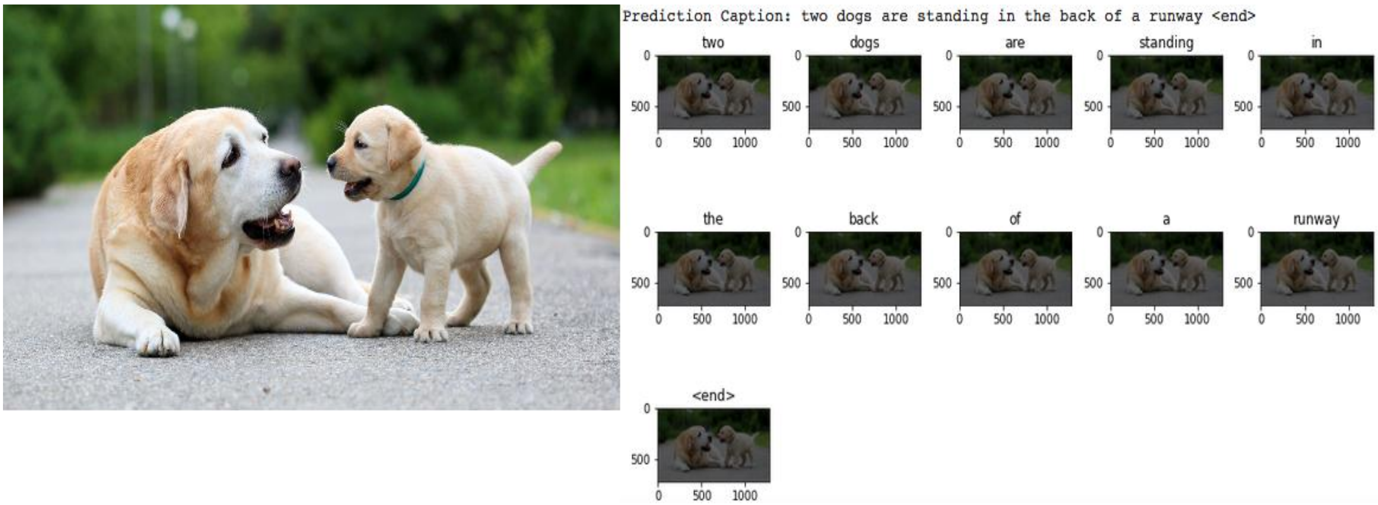
```
Prediction Caption: a woman is eating pizza <end>
```

Fig-17: Image 3 with predicted caption *"a woman is eating pizza"*



```
Prediction Caption: two people sitting on a snow covered hill with two ski poles <end>
```

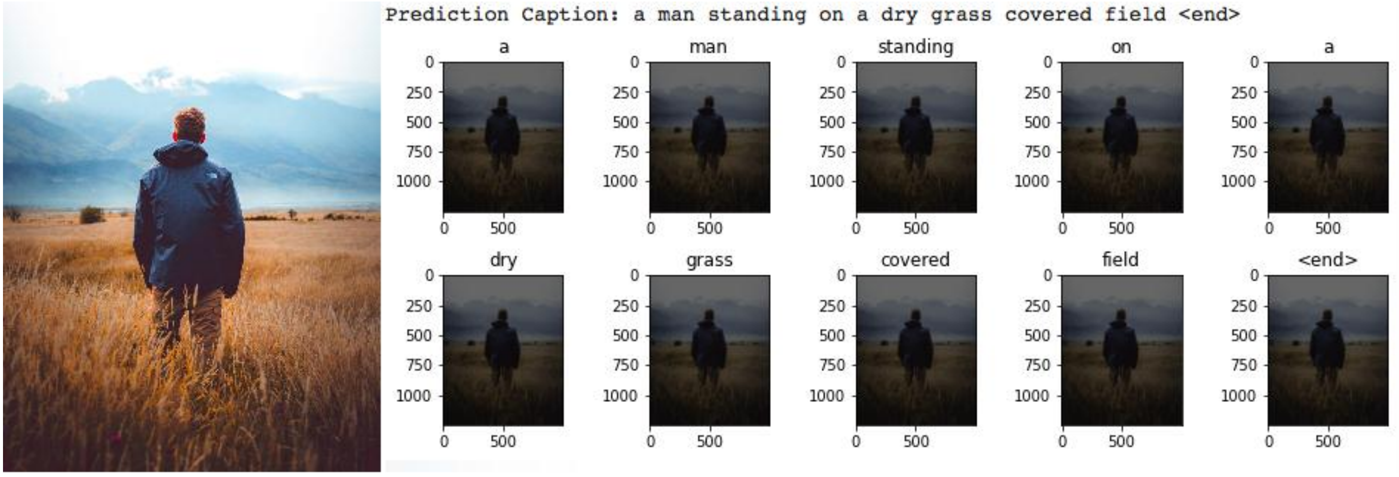
Fig-18: Image 4 with predicted caption *"two people sitting on a snow covered hill with two ski poles"*



```
Prediction Caption: two dogs are standing in the back of a runway <end>
```

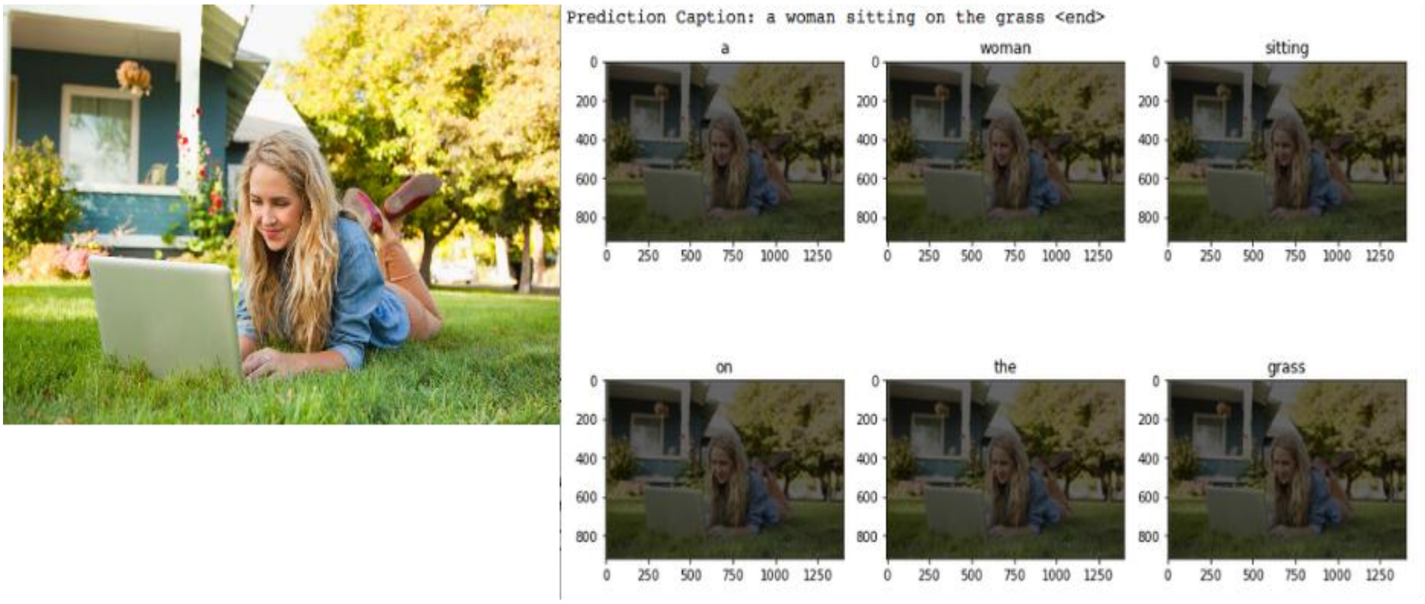
Fig-19: Image 5 with predicted caption *"two dogs are standing in the back of a runway"*

For VGG19



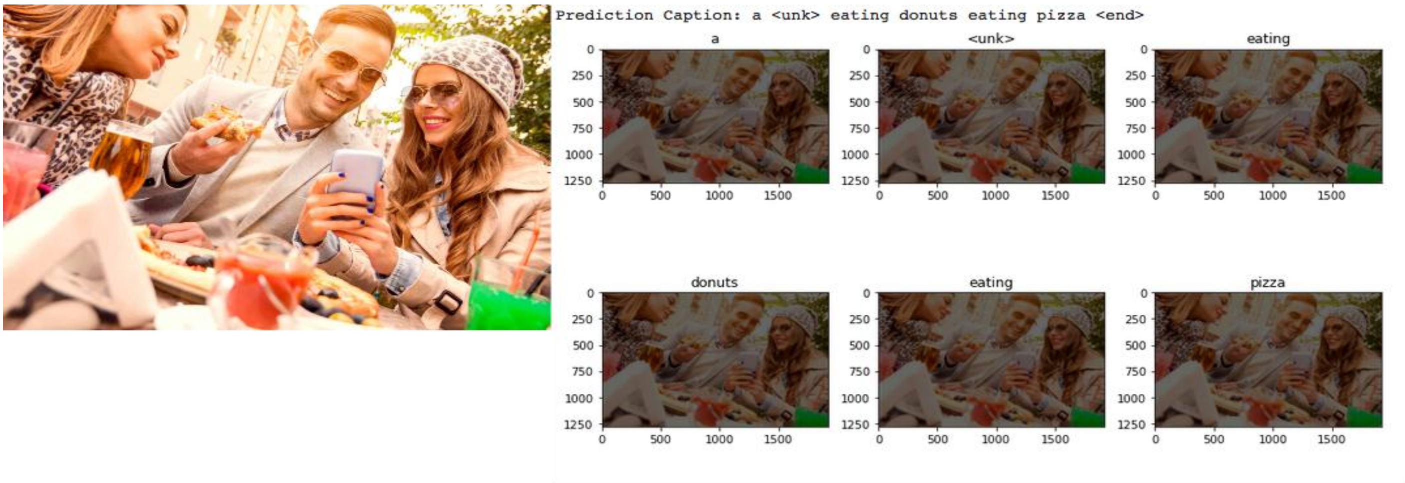
```
Prediction Caption: a man standing on a dry grass covered field <end>
```

Fig-20: Image 1 with predicted caption *"a man standing on a dry grass covered field"*



```
Prediction Caption: a woman sitting on the grass <end>
```

Fig-21: Image 2 with predicted caption *"a woman sitting on the grass"*



```
Prediction Caption: a <unk> eating donuts eating pizza <end>
```

Fig-22: Image 3 with predicted caption *"a <unk> eating donuts eating pizza"*

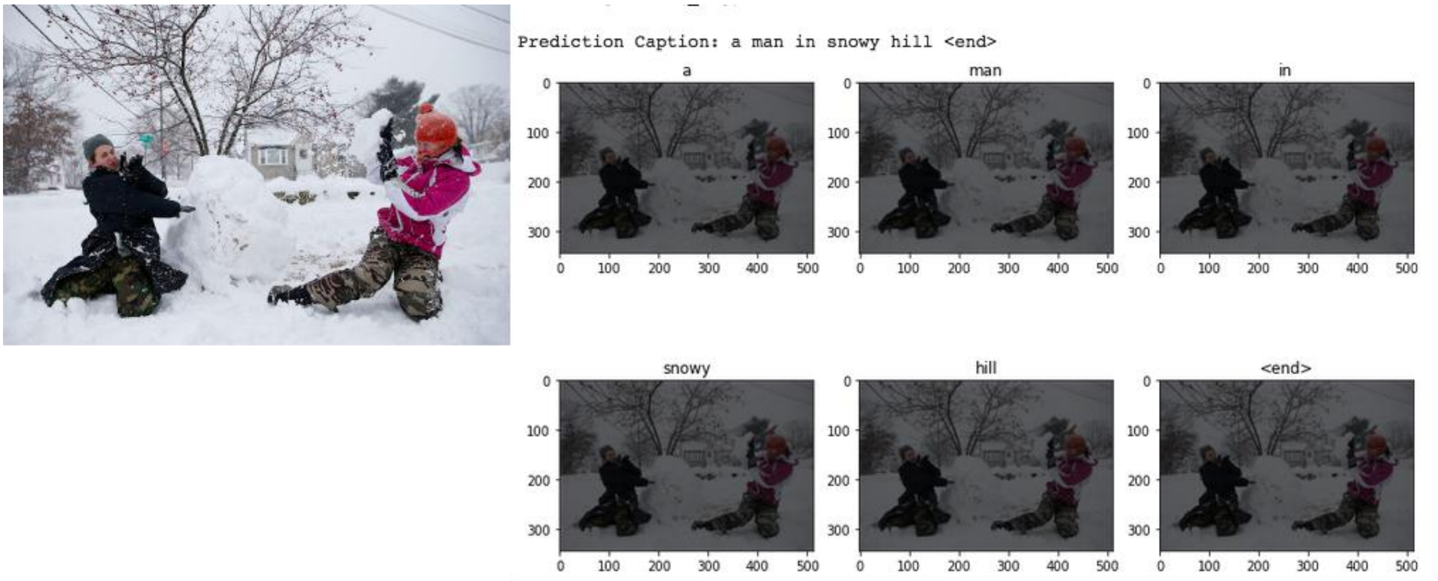


Fig-23: Image 4 with predicted caption *“a man in snowy hill”*

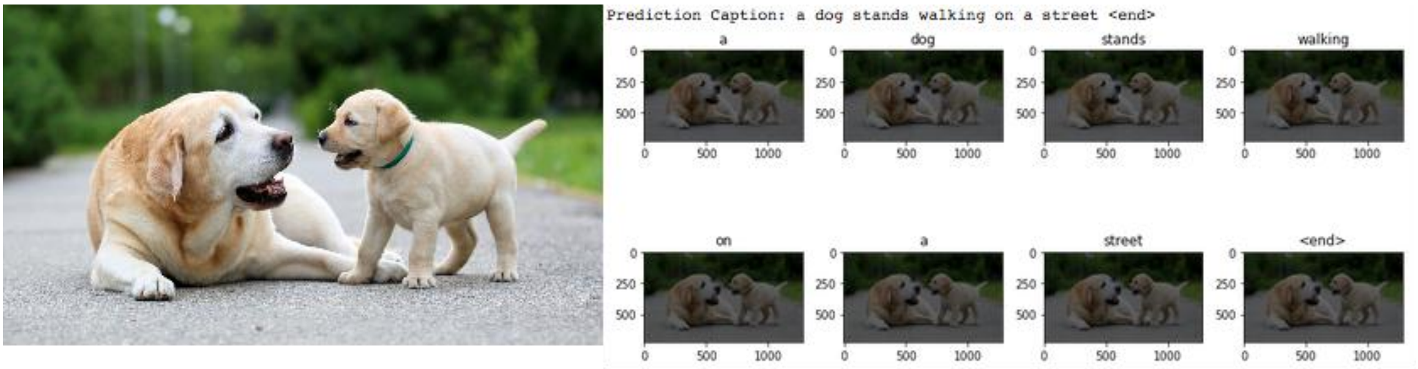


Fig-24: Image 5 with predicted caption *“a dog stands walking on a street”*

The results of these three models could be said to be apt in terms of visual attention mechanism. In a sense, the important aspects of the images were recognized. Ensemble learning is a technique where the two or more machine learning models are combined to obtain improved results. Integrating ensemble learning into the image caption models could yield much better results [11].

5. CONCLUSIONS

In this paper, we have demonstrated the results of several pre-processing models with attention mechanism in image captioning. We also reviewed different methods used in image captioning and the ways in which we can obtain improved and better results.

REFERENCES

1. 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”, 2015

2. Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel, “Multimodal Neural Language Models”, 2013.

3. MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, HAMID LAGA “Comprehensive Survey of Deep Learning for ImageCaptioning”, 2018

4. https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0

5. Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei, “Boosting Image Captioning with Attributes”,

6. https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

7. Junyoung Chung, Caglar Gulcehre KyungHyun Cho, Yoshua Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks On Sequence Modeling”, 2014.

8. https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

9. https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e

10. https://cloud.google.com/tpu/docs/inception-v3-advanced

11. Harshitha Katpally, “Ensemble Learning on Deep Neural Networks for Image Caption Generation”, 2019