2023 International Conference on Information Technologies (InfoTech-2023), Proceedings of the IEEE Conference, Rec # 58664
20-21 September 2023, Bulgaria

Describing Images using CNN and Object Features with Attention

Presented By:-

VARSHA SINGH

Research Scholar (IT)
Indian Institute of Information Technology Allahabad,
Prayagraj, U.P. India.

Contents :

- Introduction
- Problem statement
- Literature Survey
- Working mechanism flowchart
- Methodology
- Why ssd model
- Dataset
- Results
- Conclusion

Introduction :

- Image captioning : In simplest terms, it means describing an image with a sentence. It can contain anything such as objects in the image, their position, behaviour, activity etc.
- Image Captioning is for generating textual description of an image. For generating caption NLP and Computer Vision are used.
- In Image captioning,detection of the objects in image is important.

- There are so many applications of this which are :
  - CCTV surveillance
  - Image to speech conversion etc.
  - editing applications
  - image indexing
  - virtual assistants in mobile phones and computers.

Image captioning model using attention and object features to mimic human image understanding[1]

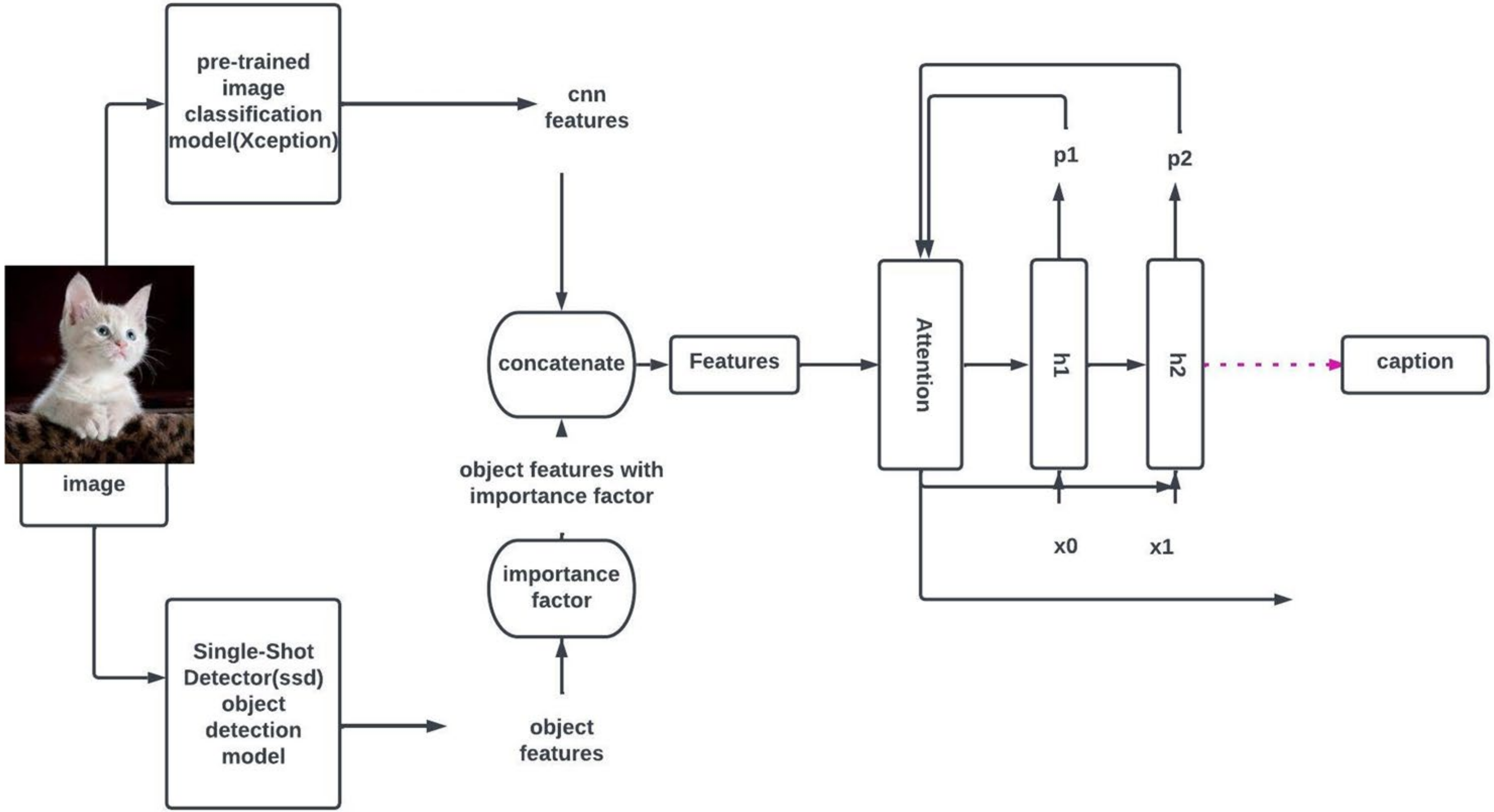
Proposed by Muhammad Abdelhadie Al-Malla1* , Assef Jafar1 and Nada Ghneim2 Al-Malla et al. Journal of Big Data (2022) 9:20 https://doi.org/10.1186/s40537-022-00571-w

- In This paper author presents an attention-based, Encoder-Decoder based architecture that makes use of convolutional features extracted from a CNN model pre-trained on ImageNet (Xception), together with object features extracted from the YOLOv4 model.
- They use the MS COCO and flikr30k dataset dataset to train and test the model.
- This paper also introduces a new positional encoding scheme for object features, the "importance factor"
- model achieves new state-of-the-art performances of **0.163** meteor score.

Problem statement :

- To generate natural textual description for an image using a combination of encoder decoder.
- There are many methodologies exist which address this problem and people have tried many experiments and have got better metrics score in general for publicly existing benchmark datasets such as COCO and Imagenets.
- Our objective is to use image identification cnn models and object detection model with transformers and see the performance.

Working mechanism flowchart :


pre-trained image classification model(Xception)
cnn features
image
concatenate
Features
object features with importance factor
importance factor
Single-Shot Detector(ssd) object detection model
object features
Attention
h1
h2
p1
p2
x0
x1
caption

Methodology:

- There are following steps in working of our method :
  - Preprocess the image
  - Extract features from the image using Xception model and the SSD which will be a sequence of numbers.
  - Concatenate the output of Xception model and SSD model, features extracted from SSD model becomes the last row of the resultant feature matrix.
  - Embedding is generated from the resultant feature matrix.
  - Embedding vector passes through lstm body and get a sequence of numbers.
  - Un-decode this vector to get the actual sentence i.e., the caption

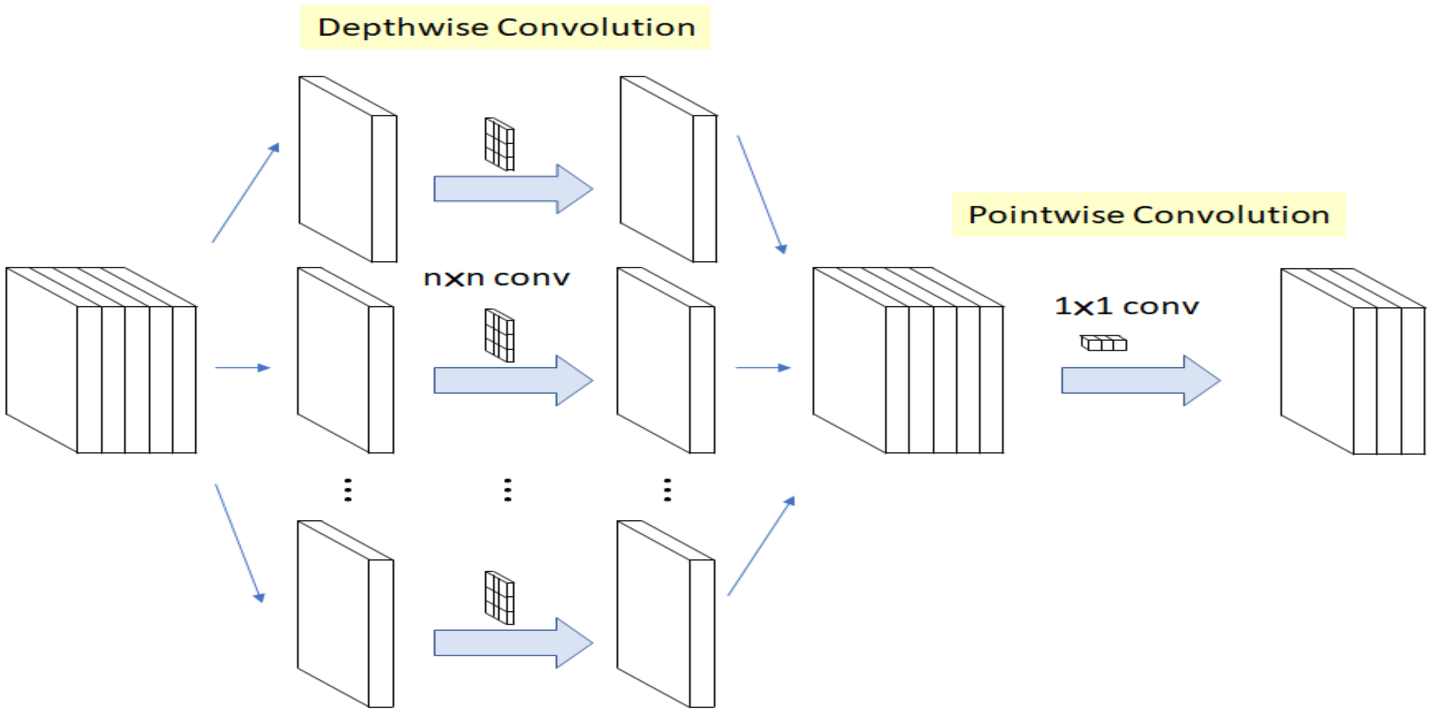
Methodology:

Let us quickly go through some terms mentioned above :

Xception:-

- Xception is a convolutional neural network that is 71 layers deep.
- This model can extract high-level visual features from images.
- Xception model has been shown to outperform other convolutional neural network architectures in terms of accuracy and efficiency on a range of image classification tasks.
- This is due in part to its use of depth-wise separable convolutions, which reduce the computational cost of the network while still maintaining high accuracy.

Methodology:

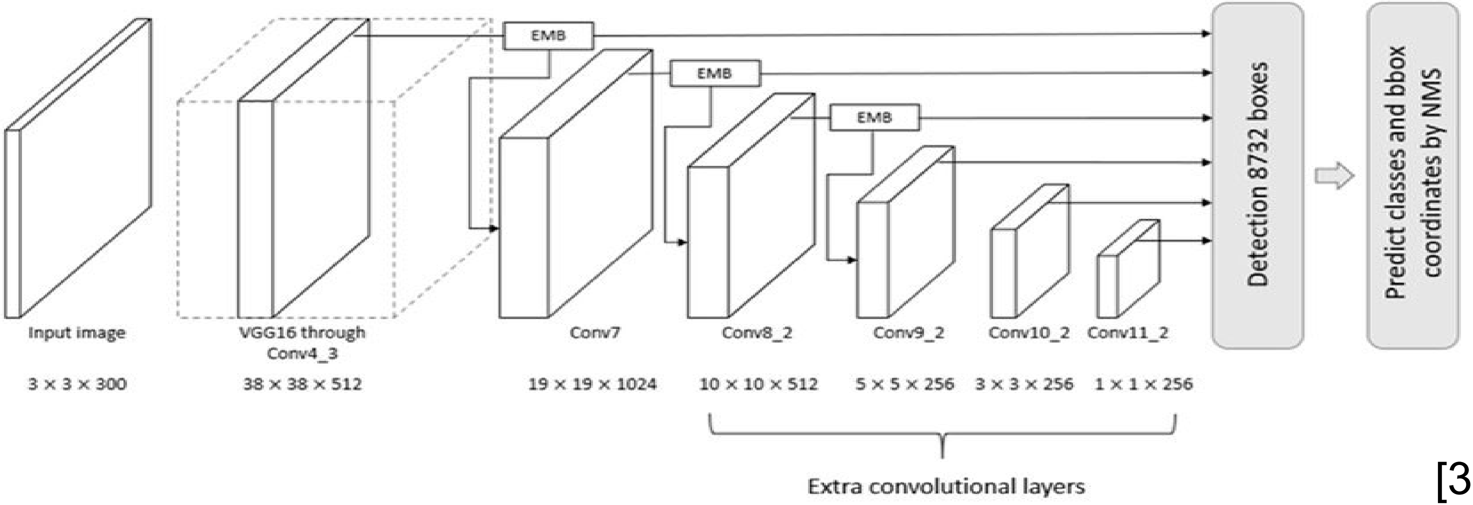


[2]

Methodology:

SSD :

- The Single Shot Detector (SSD) model is a deep learning model that's designed for object detection.
- It's notable for being fast, efficient, and accurate, making it a popular choice in a wide range of applications.
- The SSD model is based on a convolutional neural network with a predefined set of anchor boxes.
- Anchor boxes are used to detect objects at different scales and aspect ratios, allowing the model to detect

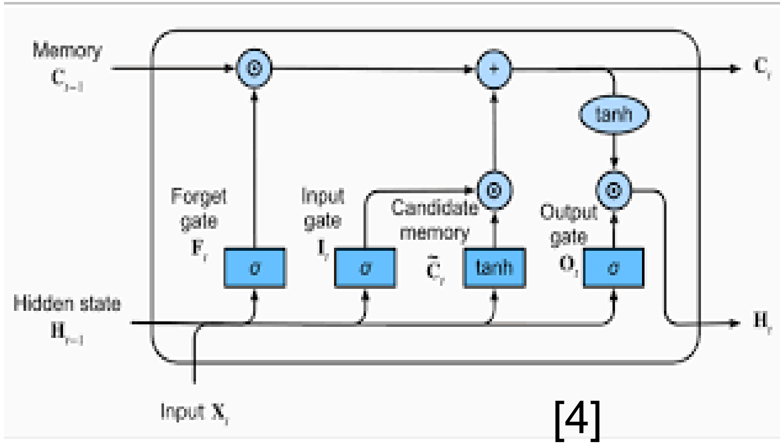


[3]

Methodology:

LSTM:

- Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that's designed to model sequential data.
- It captures long-term dependencies in data so it finds heavy utility in NLP, speech analysis and time series analysis related tasks.
- LSTM may struggle with tasks where the input data is highly noisy or contains a large amount of missing information.



[4]

Why SSD ?:

- It performs tasks of object localization and and classification in single shot or single forward pass of the network..

- It has best performance among various object detection models.

- The SSD model has shown to be **highly accurate** on a range of object detection tasks, particularly when it comes to detecting small objects.

- The SSD model can be easily adapted to different object detection tasks by adjusting the number and size of the anchor boxes. This makes it a versatile model that can be used for a wide range of applications.

- The SSD model has a relatively low memory footprint compared to other object detection models, which can be important for resource-constrained environments.

- The SSD model is designed to detect multiple objects in an image simultaneously, making it well-suited for applications where there are multiple objects of interest in the scene.

Dataset:

1. MSCOCO_Dataset

- It contains over 330,000 images and 2.5 million object instances, along with annotations for object detection, segmentation, and captioning tasks.

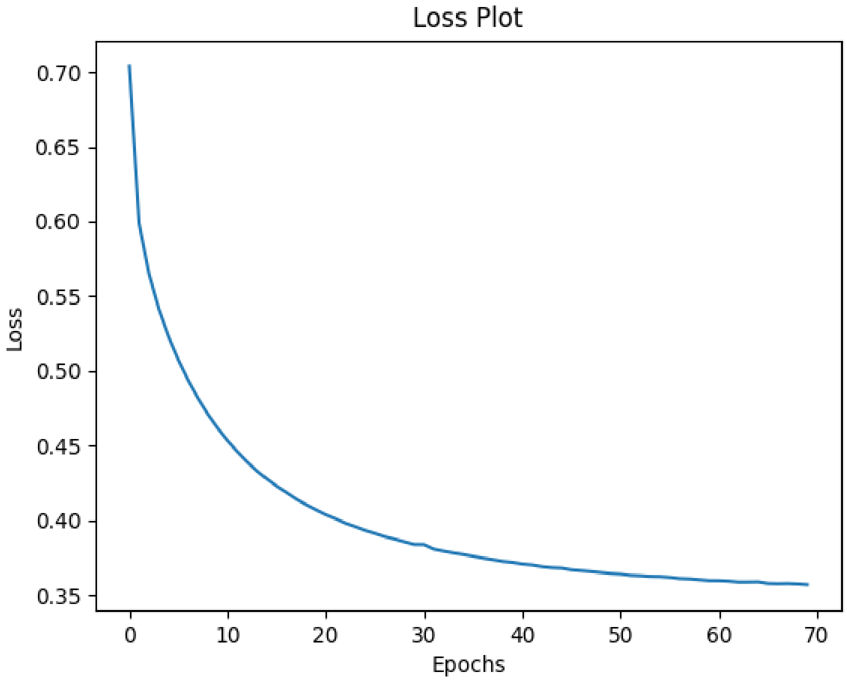
- MSCOCO is widely used for benchmark image captioning tasks.
- Training Split - 110,958
- Validation Split - 6,165
- Testing Split - 6,165

Performance Metrics:

- For our task i.e., image captioning, we have to match a sentence(ground truth) to the sentence(prediction), so we will use METEOR(Metric for Evaluation of Translation with Explicit ORdering).

Results :

Sparse categorical cross entropy loss function is used.



Results

- Baseline model solely uses Xception and we used Xception + SSD
- Incorporating SSD with Xception helped in achieving Meteor score 29.75% higher than the baseline model.
- Model's ability to relate to human evaluation significantly enhanced.
- Grammatical integrity also becomes better.
- Output caption becomes more coherent with the input image.

Meteor score of this approach is: **0.2115**

Snapshots of predicted captions :

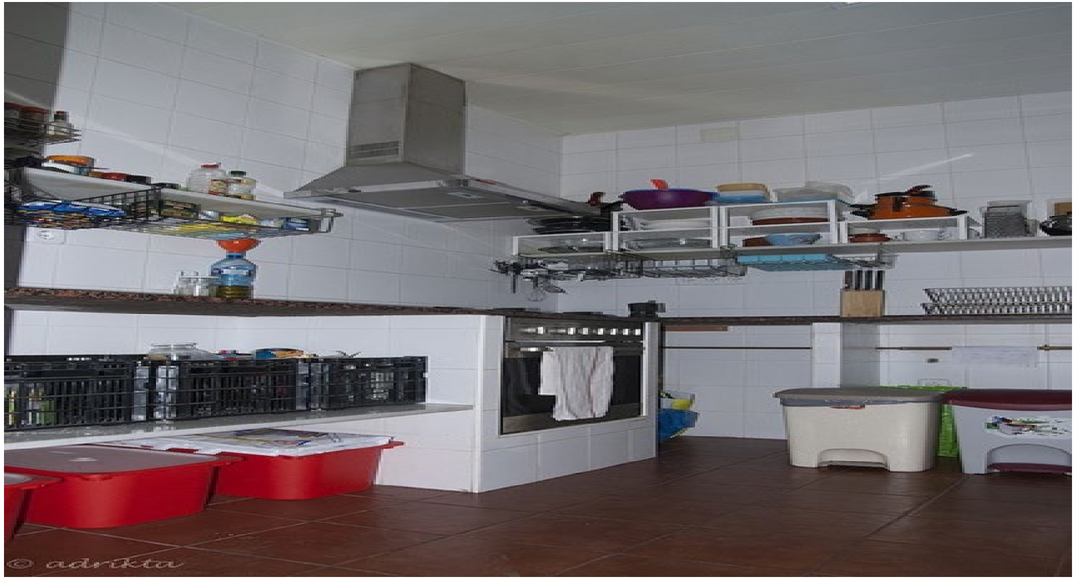


Image ID : 244328

Predicted caption : a close shot of an empty kitchen has oven interior

Original caption : a very clean and well ordered kitchen in a house

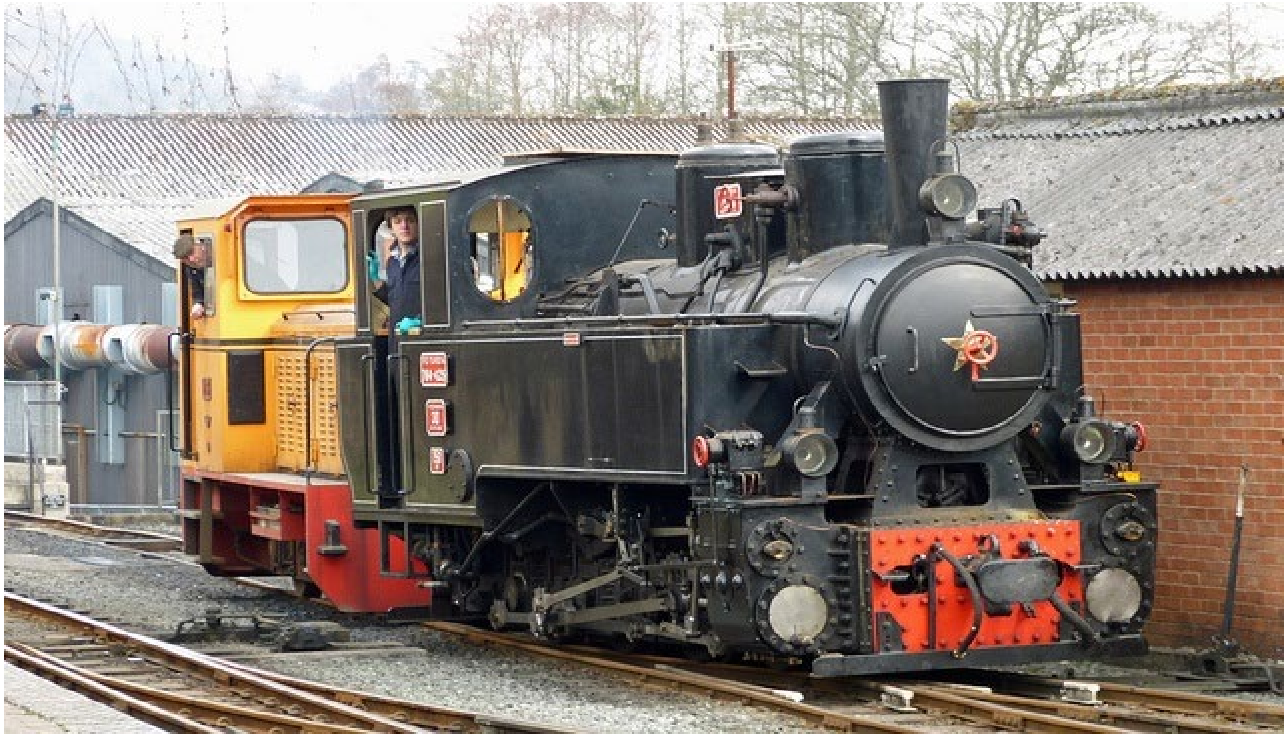


Image ID : 350059

Predicted caption : a steam engine sits atop a track under a blue sky and a sky background

Original caption : an old train is parked on railroad tracks

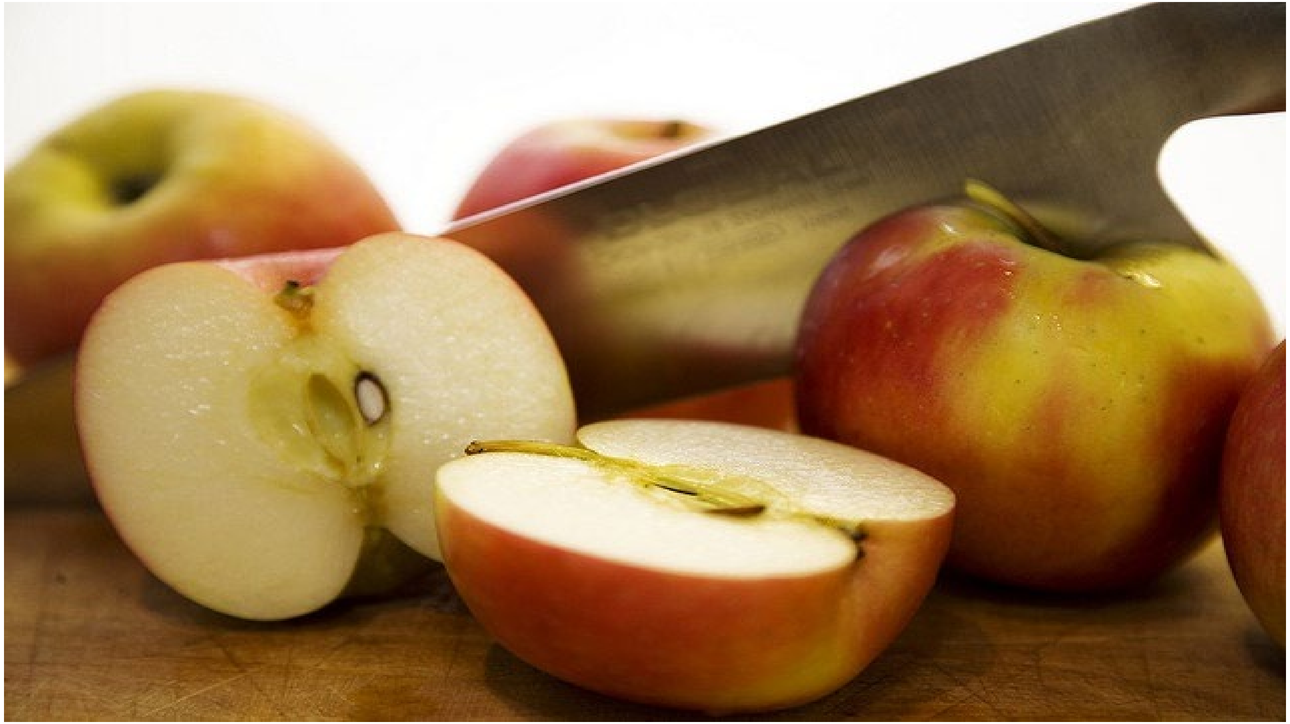


Image ID : 105697

Predicted caption : a knife with a knife and an apple and a knife and a knife

Original caption : the two halves of the apple have no blemishes or imperfections

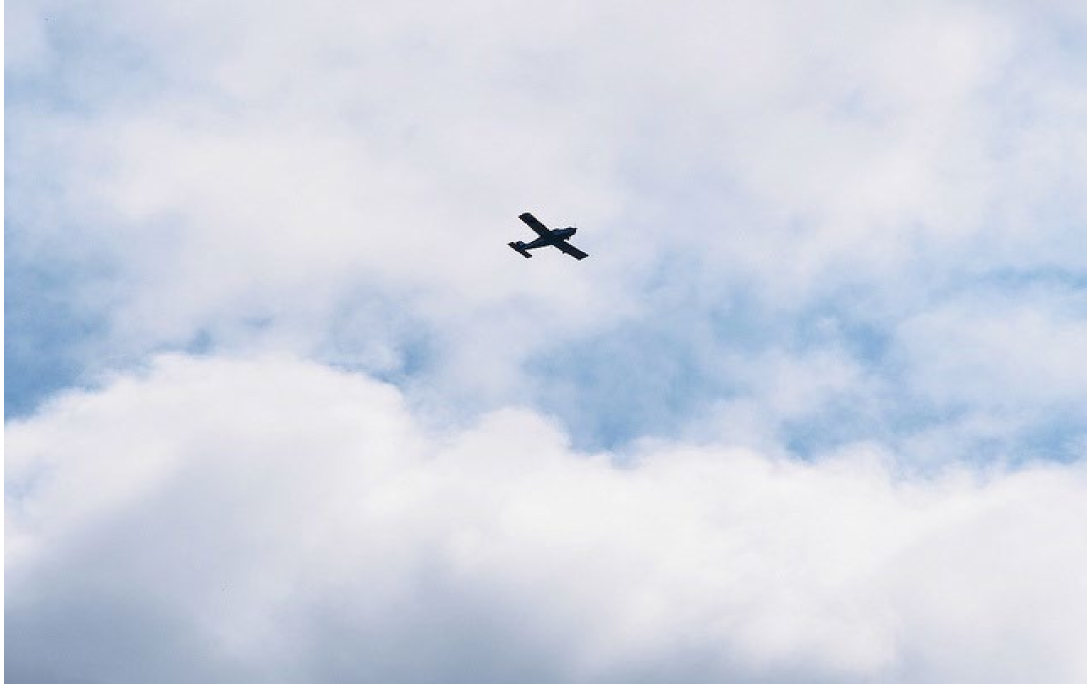


Image ID : 148299

Predicted caption : a large plane flying by in a blue sky

Original caption : a lone airplane flying in the sky with clouds around

Conclusion and future scope

- Adopted Encoder-Decoder Architecture along with attention mechanism.
- Utilized two distinct feature extraction methods (Xception and SSD).
- Using two feature extraction methods result into better and more coherent captions.
- The model can be further improved by utilizing better language generation model in the place of LSTM.
- The model can be trained and tested on other datasets like Flickr30k.

References :

1-Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), 1-16.

2-https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568

3-https://towardsdatascience.com/ssd-single-shot-detector-for-object-detection-using-multibox-1818603644ca

4-https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c

THANK YOU