International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Story Telling of a Single Image Using Redescriptions through Image Description Vision Transformer (IDVT) Algorithm

Darapu Uma[1] **Dr. M. Kamala Kumari**[2]

Submitted: 04/05/2024 **Revised**: 21/06/2024 **Accepted**: 27/06/2024

Abstract: Image Captioning is a process of transforming an input image into textual description. It uses both Computer Vision and Natural Language Processing techniques in order to generate captions. There are various image caption applications which include automation of annotation and tagging of images, self-driving cars, virtual and augmented reality applications, surveillance and security systems, object recognition and detection of images and videos. The existing techniques proposed are Bidirectional Recurrent Neural Network (BRNN), Convolution and Recurrent Neural Networks (CNN and RNN) with lack of context and appropriate meaning. The present paper proposes story telling of a single image using vision transformers. This paper narrates a story of a single image by applying a proposed algorithm named as Image Description Vision Transformer (IDVT).IDVT combines both preprocessing techniques and unsupervised algorithms of k means and mean shift to generate various descriptions of the same image and finally end up with a story.

Keywords: *Vision Transformer, Image Description, Feature Extraction, Story Telling, Redescription*

1 Introduction

Computer Vision is an interdisciplinary field that empowers to interpret and figure out visual data from the world, similar as the human visual framework. It includes a scope of strategies and innovations that permit computers to procure, process, break down, and extract images and recordings to extract significant data and go with choices in view of that data. The main aim of computer vision is an exact copy of human vision [3]. It uses digital images as input and participate various tasks like image processing, image analysis, object detection and recognition, image segmentation and biometrics etc. Computer Vision has a large number of utilizations across businesses, including medical services, car, diversion, producing, farming, security, and that's only the tip of the iceberg [3]. As innovation propels, computer vision keeps on advancing, with progressing research in regions like reasonable artificial intelligence (interpretable models), generative models, and joining vision with different modalities like language understanding for more refined simulated intelligence frameworks. Image Captioning is a technology that combines Computer Vision and Natural Language Processing (NLP) to create clear and logically pertinent text based portrayals for pictures. The objective of Image captioning to empower computers to figure out the substance of a picture and create a cognizant sentence or a short section that depicts what's going on in the picture. Image captioning has various applications such as Artificial Intelligence, Accessibility, Content Indexing, Social Media, Education and Tourism and Navigation. Be that as it may, Image captioning additionally accompanies difficulties. The created inscriptions ought to be exact, applicable, and linguistically right.

In order to produce captions that are consistent with human comprehension, the model must comprehend the image's context. Additionally, handling uncommon or unseen scenes or objects presents a challenge that necessitates robust model architectures and training data. Image captioning is an active area of research within the broader fields of computer vision and natural language processing. Image captioning models are becoming more sophisticated as AI technologies advance, able to produce high-quality captions that convey the essence of visual content.

A Vision Transformer, often abbreviated as ViT, is a type of deep learning model designed for computer vision tasks. The Vision Transformer model leverages the same transformer architecture that has been highly successful in natural language processing (NLP) tasks, such as machine translation and text generation. Image Captioning using Vision Transformers (ViTs) is an exciting application of deep learning that involves generating textual descriptions or captions for images. Vision Transformers have shown promise in various computer vision tasks, including image captioning. Vision Transformers (ViTs) are useful for image captioning due to their ability to capture rich spatial information in images and generate coherent and

[1] *Department of CSE, Pragati Engineering College, Surampalem, AP, India*
umadarapu03@gmail.com
ORCID iD: https://orcid.org/0000-0003-4037-0041

[2] *Department of CSE, Adikavi Nannaya University, Rajamahendravaram, AP, India*
mkamala@aknu.edu.in
ORCID iD: https://orcid.org/0000-0002-9358-1945

contextually relevant textual descriptions. ViTs can learn hierarchical features, from low-level details like edges and textures to high-level semantic concepts. This enables them to capture a wide range of visual information, making their representations suitable for generating descriptive captions. Vision Transformers are scalable to handle both low-resolution and high-resolution images. By splitting the image into patches and processing them independently, ViTs can effectively capture fine-grained details, making them adaptable to various image sizes and complexities .In [22] Due to their ability to model complex relationships in data, Vision Transformers often require less hand-engineering and data augmentation techniques compared to traditional convolution neural networks (CNNs). This can lead to more robust caption generation models.

Vision Transformers generate attention maps that indicate which parts of the image are most relevant when generating each word in the caption. These attention maps can provide insights into why the model generated a particular description, enhancing interpretability. Vision Transformers can be combined with text-based Transformers (like BERT or GPT) to create multimodal models. These models can jointly process both images and text, enhancing their ability to generate contextually relevant captions by incorporating knowledge from both modalities. By using sampling or beam search decoding strategies, Vision Transformers can generate diverse and creative captions for the same image. This diversity can be valuable in ensuring that the generated captions are not repetitive and appeal to a broader audience. Constructing a story using Vision Transformers (ViTs) can be a unique and creative endeavor. Vision Transformers are a type of deep learning model designed for computer vision tasks, which makes them particularly interesting for generating visual narratives.

2 Related work

Client stories have been broadly acknowledged as ancient rarities to catch the client prerequisites in agile programming improvement studied in [13]. They are short bits of texts in a semi-organized design that express prerequisites. Natural language processing (NLP) strategies offer a likely benefit in client story applications. NLP can assist system analysts manage user stories. By taking into account the investigation of NLP procedures and thorough assessment techniques is expected to acquire quality research. Similarly as with NLP research the ability to understand a sentence's context continues to be a challenge.

There are inbuilt applications that create and give a caption for an image everything are finished with the assistance of deep neural network models. It produces linguistically and semantically right sentences. In [14], it is a vital to produce image caption part of Computer Vision and Natural language processing. Image caption generators can find applications in Image segmentation as utilized by Face book and Google Photos and even all the more thus, its utilization can be stretched out to video outlines. The dataset used is Flickr8k, and the programming language used is Python3. The model consists of a vision CNN followed by a language-generating RNN. The generated captions can be used in image segmentation and can help visually impaired people. Limitations exist in object detection when using static libraries of object classes. Difficulty in establishing connections between entities within visual representations.

In [11], it discussed about image captioning with high level features. Existing CNN-RNN framework-based methods suffer from two main problems: in the training phase, all the words of captions are treated equally without considering the importance of different words; in the caption generation phase, the semantic objects or scenes might be misrecognized. In [11],it proposed a strategy in view of the encoder-decoder system, named Reference based Long Momentary Memory (R-LSTM), planning to lead the model to create a more engaging sentence for the given picture by presenting reference data. It assigns different weights to the words according to the correlation between words and images during the training phase. Finally this paper generates the quality of the sentences on Flickr 30k and MS COCO data sets .There are few limitations of this work is lack of recognition of scenes in image description and in training phase it produces equal treatment of words in captions.

Images that deal with text based data. It proposed augmenting the utilization of various modalities in an image to improve performance. It enriches the image and OCR linguistic features using pre-trained Contrastive Language-Image Pre-training (CLIP) models. We then introduce using two additional attention models in transformer architecture to strengthen the representation of the image modality [20].In this work some gaps identified such as different portrayal from references because of lower BLEU scores, Over generation issues and syntactic mistakes in produced captions and grammatical mistakes raises in generated description.

This [22] is a novel approach to generate coherent and relevant stories based on a sequence of images. A model is presented that uses a sequence encoder to process image patches and a decoder to generate human-like stories. As various evaluation metrics show, the model outperforms existing methods in terms of relevance and coherence. This paper also provides a detailed analysis of the model's performance and compares it with other state-of-the-art models. Overall, the proposed model shows promise for effectively describing a sequence of images as a story. This model is more powerful, but it lags behind in BLEU-1, BLEU-4, and CIDE revaluations. Automatic metrics may not fully reflect the accuracy of the story.

To bridge the gaps mentioned from the above papers , the proposed methodology is discussed in Section 3.

2.1 Vision Transformer

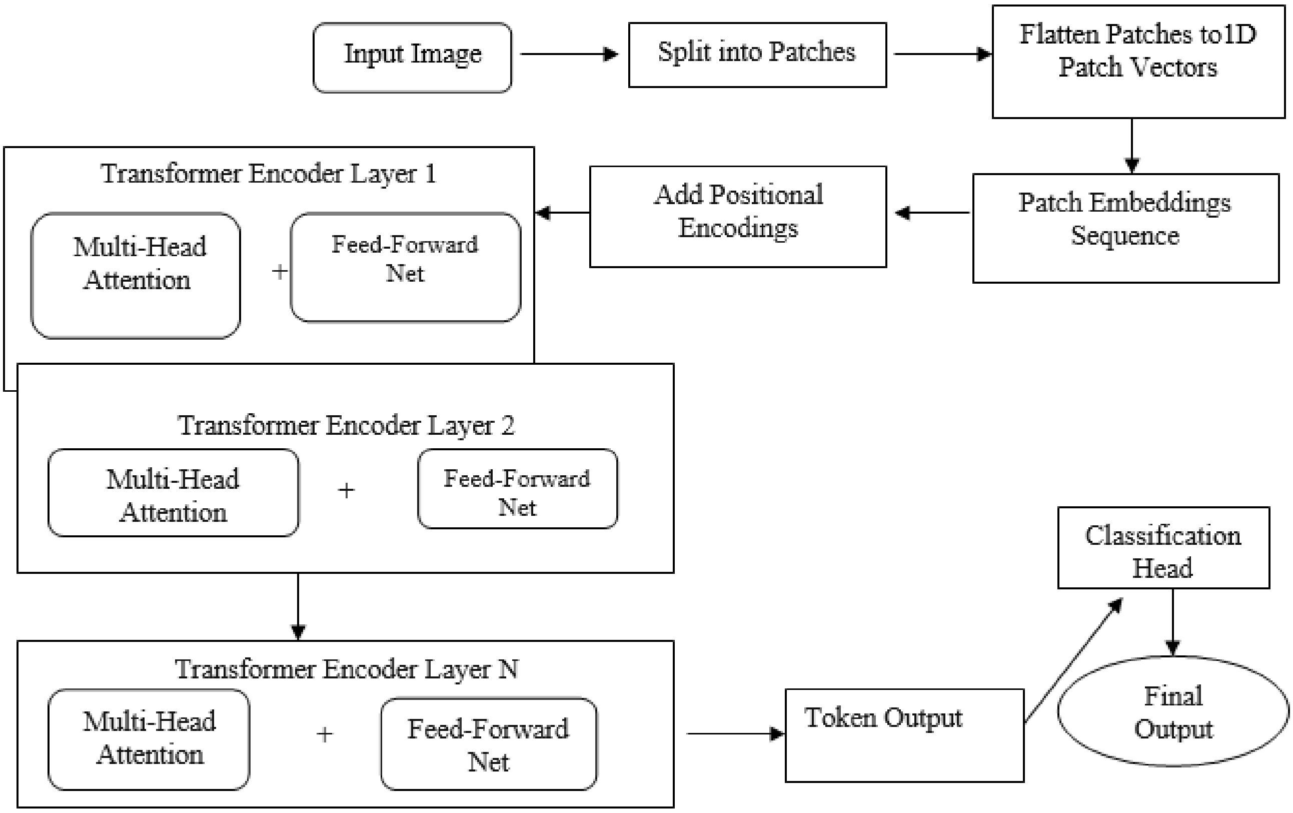
In [22], Vision Transformers, often referred to as ViTs, are a class of deep learning models that have gained popularity in the field of computer vision. They represent a significant departure from traditional Convolution Neural Networks (CNNs), which have dominated image processing tasks for many years [22]. ViTs are based on the Transformer architecture, which was originally introduced for natural language processing tasks but has since been adapted for various other domains, including computer vision [2]. The following diagram shows the architecture of Vision Transformer. An architecture called Vision Transformers (ViT) processes images with self-attention mechanisms. A number of transformer blocks make up the Vision Transformer Architecture. There are two sub-layers in each transformer block: a feed-forward layer and a self-attention layer with multiple heads. The feed-forward layer transforms the self-attention layer's output in a non-linear way, while the self-attention layer uses this relationship to calculate attention weights for each image pixel. This mechanism is extended by the multi-head attention, which lets the model focus on multiple parts of the input sequence at once. A patch embedding layer that divides the image into patches of a fixed size and maps each patch to a high-dimensional vector representation is also included in ViT [1]. These fix embedding's then taken care of into the transformer blocks for additional handling. By passing the output of the final transformer block through a classification head, which typically consists of a single fully connected layer, the ViT architecture generates a class prediction as its final output [1].ViTs are highly scalable and can handle images of different sizes without requiring architectural changes. This is in contrast to CNNs, which often need larger and more complex architectures for larger images [2]. Due to their self-attention mechanism, ViTs are effective at capturing long-range dependencies and global context in images, making them well-suited for tasks that involve recognizing complex patterns and relationships [4].

Pre-trained ViT models, similar to pre-trained language models in NLP, can be fine-tuned on specific vision tasks with relatively small amounts of task-specific data, which is a significant advantage for many practical applications [4].

The Vision Transformer plays a crucial role in image captioning by allowing automatic learning of the correspondence between image regions and their respective captions, solely with the guidance of a text description. By leveraging the power of pretrained Transformers and multi-modal learning strategies, Vision Transformer-based models have achieved state-of-the-art results in vision-language models like image captioning and vision-question answering (Kang et al., 2022). Moreover, Vision Transformers have shown great potential for various vision tasks, including image captioning (Tang et al., 2022). The importance of image captioning using Vision Transformer lies in its ability to automatically learn the relationship between image regions and their corresponding captions, solely based on text descriptions [24].This is important because it eliminates the need for manual annotation, saving time and effort in the captioning process. Furthermore, the use of Vision Transformer in image captioning enables the generation of more accurate, diverse, and coherent multi-sentence descriptions for images [24].

The importance of accurate and meaningful image captioning using Vision Transformer cannot be overstated in the field of computer vision. It allows for more advanced image understanding and analysis, opening up possibilities for applications in various domains such as autonomous vehicles, medical imaging and content-based image retrieval [24]. The integration of Vision Transformer in image captioning has shown promising results, as it effectively captures the interactions and correlations between image regions.

Fig. 1 Architecture of Vision Transformer



From the above Figure 1 represents the architecture of Vision Transformer with input images taken from the dataset. This model el utilized for image classification that implements a Transformer like structure on image patches. The image is divided into patches of a fixed or decent size, which are then linearly embedded, position embedding is added, and the resulting sequence of vectors is fed to a standard Transformer encoder. To carry out classification, the standard method of adding a learnable "classification token" to the sequence is employed [22].

Vision transformers have wide-ranging applications in well-known image recognition tasks like object detection, image segmentation, and image classification. Additionally, ViTs are utilized in generative modeling and multi-model tasks, encompassing visual grounding, Within the ViT encoder, there are multiple blocks, and Visual question answering, and visual reasoning. Each block comprises three essential processing components.

- Layer Norm
- Multi-head Attention Network (MSP)
- Multi-Layer Perceptrons (MLP)
- The Layer Norm technique ensures that the training process stays on the right track and enables the model to adjust to the variations present in the training images.
- The Multi-head Attention Network (MSP) is responsible for generating attention maps based on the embedded visual tokens provided. These attention maps assist the network in focusing on the most crucial regions within the image, such as objects.
- The MLP (Multi-Layer Perceptron) is a classification network consisting of two layers, with GELU (Gaussian Error Linear Unit) used as the activation function at the end. The final MLP block, also known as the MLP head, serves as the output of the transformer. By applying softmax to this output, classification labels can be obtained, particularly in the case of Image Classification.

3 Methodology

Fig. 2 Process Diagram of Story Construction

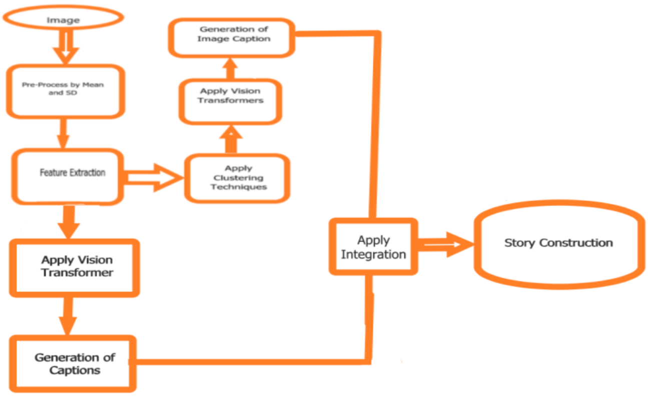


Image
Pre-Process by Mean and SD
Feature Extraction
Apply Vision Transformer
Generation of Captions
Apply Clustering Techniques
Apply Vision Transformers
Generation of Image Caption
Apply Integration
Story Construction

The above Figure 2 shows the process diagram of story construction. The input image is preprocessed by applying mean and standard deviation to remove the noise of the image. After preprocessing extract the features of the image. And apply Vision Transformer (ViT) to get description of the image. Similarly after features extraction apply clustering algorithms of K-Means and Mean Shift algorithms and output of those algorithms given to ViT to get other descriptions of the same image. Finally all those descriptions are merged to narrate a story behind them.

3.1 Image Description Vision Transformer (IDVT) Algorithm

Algorithm

Input: *Image*

Output: *Captions of the image*

1: *Read the input image*

The images taken from Flickr 8k dataset

2: *Pre-process the image with Mean and Standard Deviation*

Pre-processing can be done by applying mean and standard deviation.

Now calculating mean (μ) by

$$\mu = (1 / (M * N)) * \Sigma (i = 1 \; to \; M) \Sigma (j = 1 \; to \; N) \; I(i,j)$$

Where M represents the number of rows

N represents the number of columns

The pixel intensities are denoted by I (i, j)

Where i ranges from 1 to M and j ranges from 1 to N

Standard deviation (σ) by

$$\sigma = sqrt \; ((1 / (M * N)) * \Sigma (i = 1 \; to \; M) \Sigma (j = 1 \; to \; N) \; (I(i,j) - \mu)\text{^}2)$$

Here, sqrt () represents the square root function

Σ represents the summation symbol

3: *Estimation of 2D feature vector using Gray CO Matrix*

Using this extracting texture features like contrast, dissimilarity, homogeneity, energy, correlation, and ASM (Angular Second Moment) are used.

4:*Apply Clustering algorithm of K-Means*

Fundamental formula for K-Means

$$J(C,\mu) = \Sigma_{\mathrm{i}}\Sigma_{\mathrm{j}}\ \left|\left|x_{\mathrm{i}} - \mu_{\mathrm{j}}\right|\right|^2$$

Where

J(C, μ) is the objective function to be minimized.

C is the set of cluster assignments for each data point

(i.e., which cluster each data point belongs to).

μ is the set of cluster centroids (the mean of data points within each cluster).

Σ_i represents the sum over all data points (i) in the dataset.

5:*Apply Clustering algorithm of Mean Shift*

The Mean Shift vector, denoted as M(x),

$$M(x) = \Sigma_{\mathrm{i}}\,\phi(x_{\mathrm{i}} - x) * x_{\mathrm{i}} / \Sigma_{\mathrm{i}}\,\phi(x_{\mathrm{i}} - x)$$

Where

M(x) is the Mean Shift vector for data point x.

x is the data point for which you're calculating the Mean Shift.

x_i represents the neighbouring data points within a certain radius

(often referred to as the bandwidth or kernel width) around x.

$\phi(x_i - x)$ is the kernel function that assigns weights to neighbouring

data points based on their distance from x.

The choice of kernel function φ can vary depending on the specific application.

A common choice is the Gaussian kernel:

$$\phi(u) = (1 / (2\pi\sigma\text{^}2\,)) * e\text{^}(-|(|u|)|\text{^}2/\,(2\sigma\text{^}2\,)\,)$$

Where

u is the vector representing the distance between data points x and x_i.

σ is the bandwidth parameter

6: *Apply Vision Transformers*

ViT takes an image as input, which is divided into non-overlapping patches.

Each patch is then linearly embedded into a lower-dimensional vector.

The patch size (p) can be calculated as:

$$p = sqrt((W * H) / P\,)$$

Where

image has dimensions of WxH, divide it into P patches

The positional encoding for a patch at position (x, y) is calculated as:

$$PE(x,y) = [sin(x/10000\text{^}i/d), cos(x/10000\text{^}i/d)]$$

Where i is the dimension index, and d is the dimension of the positional encoding

7: *Generate Caption of the image using Vision Transformer (ViT)*

8: *Integrate the outputs of Step (4), (5) and (6)*

From the above algorithm, input images have been taken from the Flickr 8k dataset and prep processing of those images by applying mean and standard deviation and then extract the feature vectors of Contrast, Energy, ASM, Dissimilarity, Homogeneity and Correlation. Applying unsupervised algorithms of K means and Mean shift clustering algorithms to enhance the quality of the image like intensity of the image. The image s after the clustering algorithms has to be applied to a vision transformer to predict the description of the image. The vision transformer uses pre-trained model of vision encoder and

decoder and assist to translate the image into text manner. Finally getting different descriptions of the same image those descriptions will be integrated, to get a story behind them.

4 Results & Discussions

4.1 Read the input image

In Fig 5 images have been taken from the Flickr 8k dataset. These images have been preprocessed by applying the mean and standard deviation to enhance the quality of the images and remove the noise from the data. After applying the gray co matrix to extract the features of the given images, those images can be shown below.

Fig. 5 Sample Images From Flickr8k Dataset

In the fields of computer vision and Natural Language Processing (NLP), the Flickr 8K dataset is widely used. It is used for several research and development purposes, such as image captioning, multimodal AI, evaluation of language models, human-AI interaction, and content generation. There are some datasets that support captioning of an image, such as the Microsoft COCO dataset [5]. It makes use of these datasets, which consist of Flickr images. 8K and 30K. The majority of image captioning research There are 31,000 images in Flickr 30K and 8,000 in Flickr 8K, each with multiple captions. When compared to COCO, these datasets have a relatively smaller scale. Image captioning. It contains over 200,000 images, each paired with multiple human-generated captions and Flickr When compared to some of the more recent datasets, this is one of the most popular and widely used datasets because the dataset has a relatively small number of images and captions, which may limit its diversity and ability to apply to situations that occur in the real world. It is often used as a starting point for more comprehensive image captioning models or combined with other larger datasets.

4.2 Pre-process the image with Mean and Standard Deviation

In image processing and analysis, the concepts of mean and standard deviation are frequently applied to images for a variety of purposes. An image's mean is the average intensity value across all of its pixels. It gives an indication of the image's overall brightness or darkness. The variation or spread of pixel intensities around the mean is measured by an image's standard deviation. Mean and standard deviation are often used in image preprocessing, particularly in the context of normalization. Normalization is a common image preprocessing technique used to enhance the training process of machine learning models, especially deep neural networks. It involves transforming the pixel values of an image to a specific range or distribution to

ensure stable and efficient learning .Now preprocessed the images with mean values by applying mean on the image with respect to the given mentioned formula.

The mean of an image is calculated by summing up all the pixel intensities in the image and dividing it by the total number of pixels

The formula for calculating the mean is

$$\mu = (1 / (M * N)) * \Sigma\,(i = 1 \text{ to } M)\, \Sigma(j = 1 \text{ to } N)\, I(i, j) \quad (1)$$

Where

Image with dimensions M x N

Where M represents the number of rows

N represents the number of columns

The pixel intensities are denoted by I (i, j)

Where i range from 1 to M and j ranges from 1 to N

The above Eq (1) represents mean formula; this can be applied to an image to enhance the pixel intensity values of the image.

The standard deviation of an image quantifies the variety or spread of pixel values within the image. It estimates how much the individual pixel values deviate from the mean. A higher standard deviation demonstrates that the pixel values are more spread out and that the image may have more contrast or texture, while a lower standard deviation suggests a smoother, less varied image.

Now preprocessed the images with standard deviation values by applying standard deviation on that image using the below formula. The standard deviation of an image measures the spread or dispersion of pixel intensities around the mean. It is calculated by taking the square root of the average of the squared differences between each pixel intensity and the mean.

The formula for calculating the standard deviation is

$$\sigma = \mathrm{sqrt}\left((1 / (M * N)) * \Sigma\,(i = 1 \text{ to } M)\,\Sigma\,(j = 1 \text{ to } N)\,(I(i,j) - \mu)^{\wedge}2\right) \quad (2)$$

Here, sqrt () represents the square root function

Σ represents the summation symbol.

The above Eq (2) represents standard deviation formula, this can be applied to an image to enhance the pixel intensity values of the image. These equations (1) and (2) allow you to compute the mean and standard deviation of an image, providing valuable statistical information about its intensity distribution and variation.

The mean and standard deviation values of the input images are given below

Table 3 Mean and Standard Deviation values of the input images

Image	Mean	Standard Deviation
(a)	(93.54962666666667, 78.16518933333334, 56.600704)	(62.49954365,54.14677419, 52.23228618)
(b)	(120.20346779,120.2970084, 119.12667227)	(83.02585758,84.88645875, 88.57879592)
(c)	(150.51322523,144.72036637, 128.17151351)	(57.98774717,47.83739397, 48.92862971)
(d)	(142.5981021,151.06956757, 123.26628829)	(68.72702135,58.73726941, 83.72440205)
(e)	(132.23905988,116.08401796, 134.14972455)	(89.47336569,94.53279329, 96.89571461)
(f)	(110.882464,132.66916267, 162.113552)	(62.43859762,61.55293681, 61.9138193)
(g)	(89.07324074,104.86543827, 67.01795679)	(49.30455204,51.25592876, 42.31866024)
(h)	(140.84365766,127.93637237, 84.98340541)	(52.01379762,51.42738476, 38.86467885)
(i)	(130.8221848,141.29634908, 104.54328131)	(60.61726575,66.9022559, 55.81735575)

The above Table shows the feature vectors of the sample images taken from the flickr8k dataset.

4.3 Estimation of 2D feature vector using Gray CO Matrix

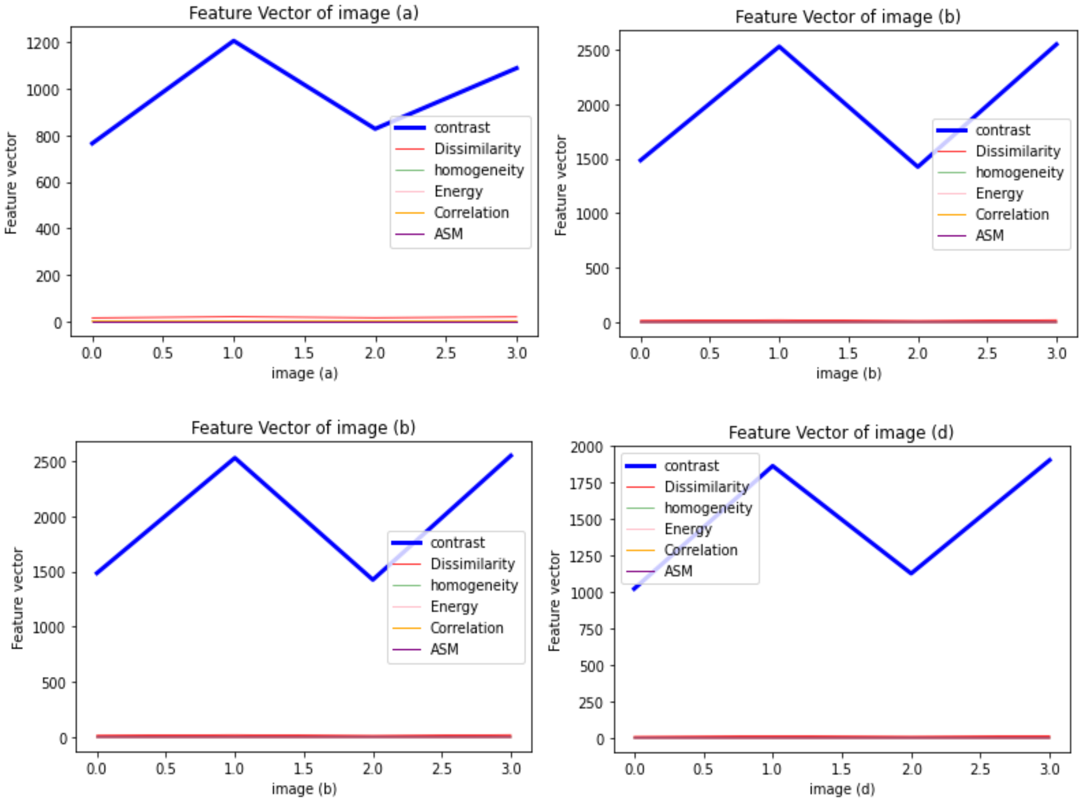
From an image's co-occurrence matrix, texture features like contrast, dissimilarity, homogeneity, energy, correlation, and ASM (Angular Second Moment) are frequently used. An image's texture properties are revealed by these features. Contrast defines the image's local intensity variations are measured by contrast. It measures the intensity differences between a pixel and its adjoining pixels. Greater differences in the intensities of adjacent pixels are indicated by contrast values that are higher. Dissimilarity defines the average intensity difference between a pixel and its neighbors is measured by dissimilarity. It records how much local variation is available in an image. A greater variation in the intensities of adjacent pixels is indicated by dissimilarity values that are higher. Homogeneity assesses the consistency of connecting pixel intensities. It measures how close the intensities of the pixels are to each other. The presence of similar intensities between adjacent pixels is shown by higher homogeneity values. Energy characterizes the amount of squares or consistency, gauges the amount of squared components in the co-occurrence matrix. It gives a number to the texture's overall uniformity or smoothness. A texture that is more uniform or homogeneous is shown by higher energy values .Correlation measures the linear dependency between the pixel intensities and their neighboring pixels. The linear relationship between the pixel values is quantified by it. While a negative correlation indicates that neighboring pixels are changing in opposite directions, a positive correlation indicates that neighboring pixels are changing together.ASM(Angular Second Moment) ascertains the amount of squared probabilities in the co-occurrence matrix. The overall uniformity or orderliness of the texture is quantified. A more uniform or organized texture is shown by ASM values that are higher.The gray-level co-occurrence matrix (GLCM), a statistical representation of the spatial relationships between pixel intensities in an image, is frequently used to calculate these texture features. At a predetermined distance and orientation, the GLCM records the occurrence frequencies of specific pixel intensity pairs. By extracting these texture features, ne can describe and break down the texture properties of an image These features are used in image analysis, pattern recognition, and computer vision, all of which rely heavily on texture information, among other areas. The below Table shows the feature vectors of the sample images taken from the flickr8k dataset.

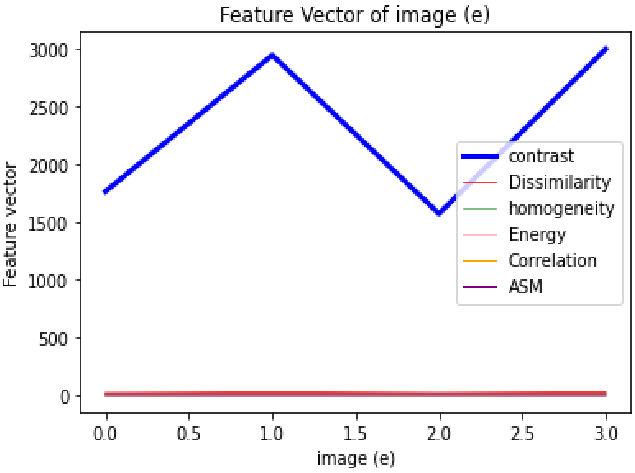
Table 4 Feature Vectors of the above Sample input Flicker 8k Dataset images

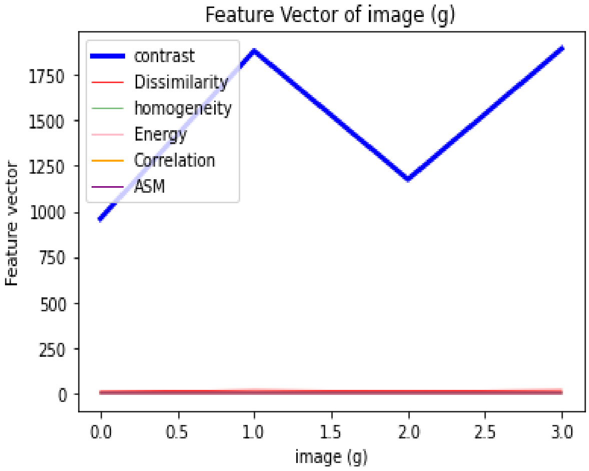
Image	**Contrast**	**Dissimilarity**	**Homogeneity**	**Energy**	**Correlation**	**ASM**
(a)	[765.38879144 1206.49814602 827.9671984 1088.91297568]	[15.49683422 20.21728484 16.0807161 19.56323878]	[0.13382814 0.10754146 0.14165092 0.10643864]	[0.01273495 0.01123591 0.01285462 0.01129054]	[0.87221328 0.79858166 0.86214187 0.81821451]	[0.00016218 0.00012625 0.00016524 0.00012748]
(b)	[1484.5347309 2527.78167795 1424.2999861 2548.1454009]	[14.95619881 20.50659171 12.88158992 20.67812714]	[0.37564743 0.32219004 0.36487353 0.32105583]	[0.20636557 0.19736091 0.20480561 0.19744562]	[0.91449632 0.85418577 0.91793234 0.8530111]	[0.04258675 0.03895133 0.04194534 0.03898477]
(c)	[1121.87973472 2242.96455821 1389.91121851 2201.60652732]	[11.61656882 20.41718059 15.18442079 20.17549528]	[0.3544173 0.27907662 0.30382572 0.27890475]	[0.20278424 0.19413201 0.2009754 0.19423154]	[0.86746238 0.73399463 0.83560141 0.73889951]	[0.04112145 0.03768724 0.04039111 0.03772589]
(d)	[1022.87647315 1865.67614474 1127.78899702 1905.73124488]	[8.9552294 13.83803962 9.91670477 14.34256399]	[0.45048013 0.36687601 0.39713759 0.36604953]	[0.20402303 0.19514705 0.20210412 0.19521619]	[0.90758535 0.83111959 0.89804695 0.82749381]	[0.0416254 0.03808237 0.04084608 0.03810936]
(e)	[1762.55948068 2941.28130954 1569.30474857 2993.96769444]	[14.03968254 22.21111207 14.43557661 22.60544777]	[0.47687699 0.41749779 0.47558076 0.42050198]	[0.3042531 0.29493609 0.30802384 0.29471805]	[0.90706666 0.84491757 0.91725551 0.84213962]	[0.09256995 0.0869873 0.09487869 0.08685873]
(f)	[1230.04687123	[9.03608819	[0.59106479	[0.2138089	[0.89524082	[0.04571424

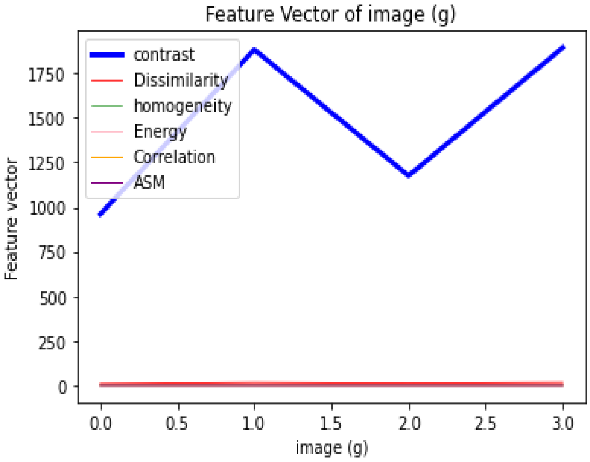
	2294.37895833 1340.43996137 2233.80394531]	15.66410346 10.48902572 15.15671054]	0.52612068 0.56067484 0.51917183]	0.20479785 0.21266507 0.20483201]	0.80399391 0.88575675 0.80916876]	0.04194216 0.04522643 0.04195615]
(g)	[959.15533486 1881.92249185 1175.62855777 1894.02119863]	[9.7827434 16.58823529 12.486417 16.86290134]	[0.34792606 0.28448334 0.31223461 0.28537197]	[0.20701734 0.1984528 0.20511261 0.19853689]	[0.92309862 0.8481299 0.90554259 0.84715354]	[0.04285618 0.03938351 0.04207118 0.0394169]
(h)	[1113.48597521 1977.61283633 1116.78984608 1918.5838563 6]	[12.26714503 17.68114392 11.56930747 17.19067838]	[0.32355953 0.28237346 0.32215141 0.28430115]	[0.20252849 0.19404079 0.20092753 0.19414328]	[0.8838865 0.79271141 0.88335629 0.79889869]	[0.04101779 0.03765183 0.04037187 0.03769161]
(i)	[1389.90384494 2457.57359407 1376.96012774 2427.71825844]	[13.61392525 21.31232838 14.94472902 21.19686989]	[0.3442402 0.29954816 0.32999075 0.29993183]	[0.241509 0.23308963 0.24266069 0.23320295]	[0.88131087 0.78968523 0.88241619 0.7922402]	[0.0583266 0.05433078 0.05888421 0.05438361]

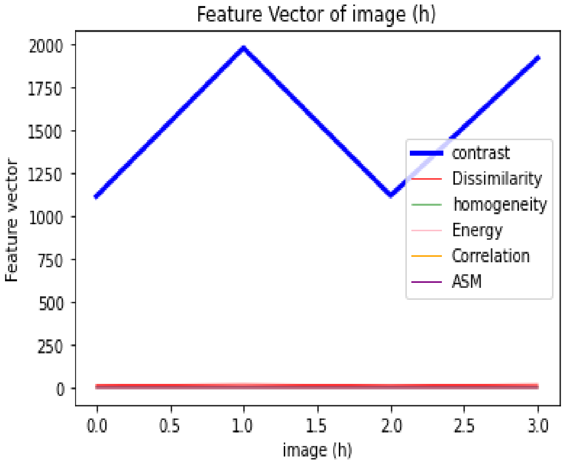
Fig .6 Plot diagrams of Feature Vector of the input images

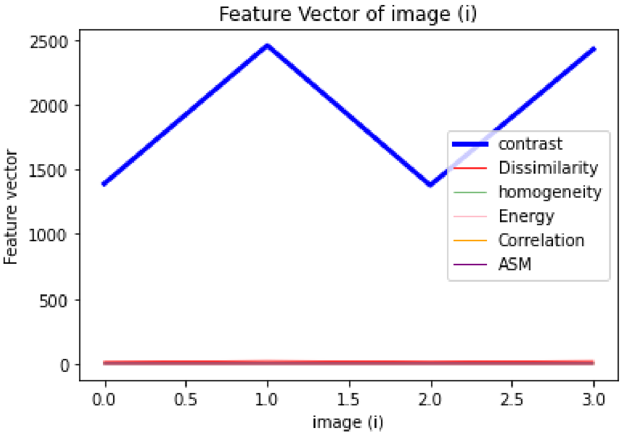












The above diagram shows the visualization form of the feature vectors of each image. These features have extracted from the GrayCoMatrix. Each of the above images indicates the various features of the images from (a) to (i) and all those features are differentiated with various colors of blue, purple, orange etc. Each image shows the feature vectors like contrast, brightness, dissimilarity a single image taken from Flickr 8k dataset. The above plot diagrams are unique according to their features.

4.4 Apply Clustering algorithm of K-Means

Clustering algorithms can be used in image captioning for several purposes, mainly to improve the quality and variety of generated image captions. Here are some of the key purposes and benefits of using clustering algorithms in image captioning: Semantic Grouping of Concepts, Diversity in Captions, Improved Caption Generation, and Reduction of Caption Length [5]. This paper performs clustering based image segmentation to get valid descriptions of the image. Semantic Grouping of Concepts: Within an image, clustering algorithms can assist in grouping like scenes, concepts, or objects. Clustering, for instance, has the ability to assemble all of the dog instances together in an image with multiple dogs. This semantic gathering can prompt more intelligible and logically exact subtitles.

4.4.1 Diversity in Captions

By recognizing and clustering various objects or elements in an image, clustering algorithms are able to guarantee that the produced captions are diverse and cover multiple aspects of the image. By doing this, descriptions won't have to be the same every time.

4.4.2 Improved Caption Generation

To provide a more comprehensive description of the image, captions can be written for groups of objects rather than for each individual object.

4.4.3 Reduction of Caption Length

Grouping can prompt more compact and valuable Iterative clustering with the K-Means algorithm divides a dataset into K clusters, with each data point belonging to

the cluster with the closest mean. The sum of squared distances between each data point and the mean of its assigned cluster is minimized by the algorithm. The K-Means algorithm's fundamental formula is as follows:

Objective Function (to be minimized):

$$J(C, \mu) = \Sigma_i \Sigma_j \left|\left|x_i - \mu_j\right|\right|^2 \quad (3)$$

Where:

$J(C, \mu)$ is the objective function to be minimized.

C is the set of cluster assignments for each data point (i.e., which cluster each data point belongs to).

μ is the set of cluster centroids (the mean of data points within each cluster).

Σ_i represents the sum over all data points (i) in the dataset.

Σ_j represents the sum over all clusters (j).

4.5 Apply Clustering algorithm of Mean Shift

The Mean Shift algorithm is a non-parametric method for clustering that is used to find modes or peaks in the probability density function of a dataset. It's not unexpected applied in image segmentation and data clustering tasks [5]. The Mean Shift algorithm's basic formula is as follows

Mean Shift Vector:

The Mean Shift vector, denoted as M(x), represents the direction and magnitude of the shift for a data point x during each iteration [8]. It is calculated as follows:

$$M(x) = \Sigma_i \phi(x_i - x) * x_i / \Sigma_i \phi(x_i - x) \quad (4)$$

Where:

M(x) is the Mean Shift vector for data point x.

x is the data point for which you're calculating the Mean inscriptions. The caption can focus on the clusters and their relationships rather than describing each individual object in detail, resulting in captions that are more restricted and easier to read.

K-Means and Mean Shift clustering, are instances of common clustering algorithms utilized in image captioning. With the help of these algorithms, images can be preprocessed, salient regions can be identified, and similar visual elements can be grouped together, resulting in captions for images that are more exact and relevant to the context [9]. However, careful design is often required when combining clustering with caption generation to ensure that generated captions are informative and coherent [5].

Shift.

x_i represents the neighbouring data points within a certain radius (often referred to as the bandwidth or kernel width)

around x.

$\phi(x_i - x)$ is the kernel function that assigns weights to neighbouring data points based on their distance from x.

Kernel Function:

The choice of kernel function ϕ can vary depending on the specific application. A common choice is the Gaussian kernel

$$\phi(u) = (1 / (2\pi\sigma\hat{}2)) * e\hat{}(-|(|u|)|\hat{}2/ (2\sigma\hat{}2)) (5)$$

Where:

u is the vector representing the distance between data points x and x_i.

σ is the bandwidth parameter that controls the size of the kernel and influences the clustering result.

Iteration:The Mean Shift algorithm iteratively updates the position of each data point in the direction of the Mean Shift vector:

$$x\_new = x + M(x) \quad (6)$$

This process is repeated until convergence, which occurs when the Mean Shift vector becomes very close to zero (i.e., the data point reaches a mode or peak in the data distribution).The Mean Shift algorithm calculates the Mean Shift vector for each data point, which points towards the mode of the underlying data distribution [9]. It then updates the position of each data point in the direction of this vector until convergence. As a result, data points within the same mode are attracted to the same position, effectively grouping them into clusters. The algorithm is adaptive and doesn't require specifying the number of clusters before hand, making it suitable for applications where the number of clusters is not known in advance [5].

4.6 Apply Vision Transformers

Vision Transformers are a sort of transformers that perform visual-related errands that incorporate images [8]. First introducing the transformer library and afterward fabricate the model prior to utilizing our model to produce captions of images. In fact, we are making use of the Hugging Face library's vit-gpt2-image-captioning model, which has been trained to caption images[1].

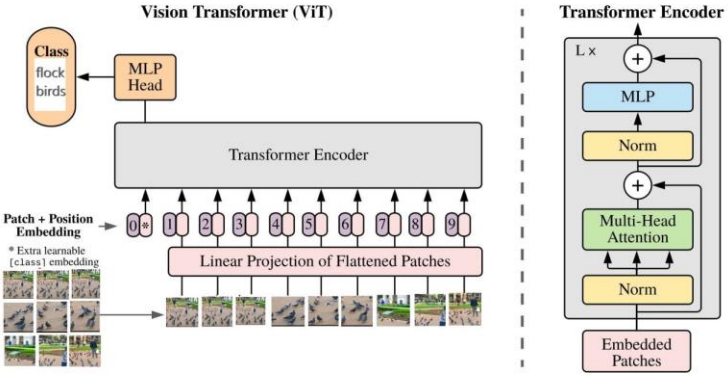


Fig .7 Image Captioning Using Vision Transformer

The above figure shows that image captioning model using vision transformers. It has three pre-trained models.

- Vision Encoder Decoder Model
- GPT2Tokenizer
- Fast ViT Image Processor.

Vision Encoder Decoder Model:

This helps in doing an image-to-text generation with any pre-prepared vision model utilizing a Transformers (as the encoder) like ViT (which we utilized here) or BEiT kind of models which uses self-regulated pre-preparing of Vision Transformers (ViTs) to outperform supervised pre-training alongside any pre-trained language model as the decoder such as GPT2 . As a result, we use Vision Encoder Decoder as an application for image captioning in this technique encodes the image and then generates captions using a language model

GPT2TokenizerFast:

Using the Hugging Face tokenizers library, this creates a GPT-2 tokenizer. The transformers are loaded with the tokenizers library. The tokenizer as of now has the skills necessary to perform all of the tasks required for captioning.

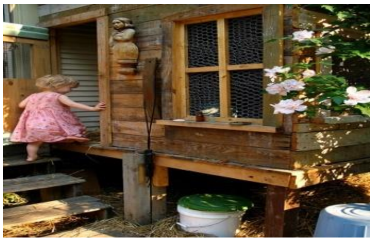
ViT Image Processor:

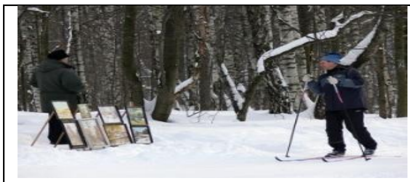
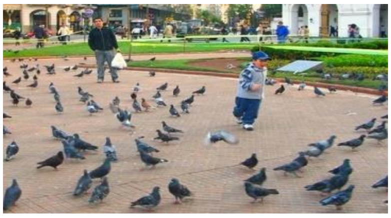
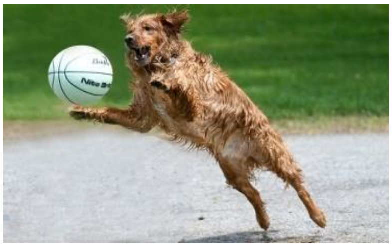
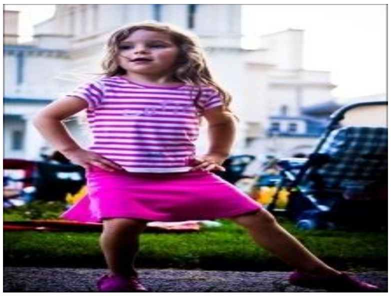
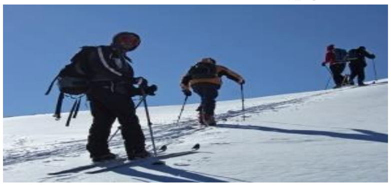
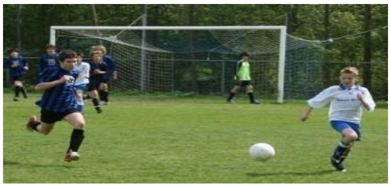
It supportss in the design of a ViT image processor. After that load the images and generates captions of the images.

4.7 Generate Caption of the image using Vision Transformer (ViT)

Table 5 below shows the results of the IDVT proposed algorithm. The proposed algorithm generates the captions of the images using the vision transformer result and also generates different captions by applying the unsupervised algorithms of K-Means and the Mean Shift algorithm.

Table 5 IDVT Algorithm Generates Captions of images

Image	Transformer Result	K-Mean Result	Mean Shift Result
1000268201_693b08cb0e.jpg 	['a little girl standing next to a wooden bench']	['a pile of junk sitting on top of a wooden structure']	['a little girl standing in front of a window']
101669240_b2d3e7f17b.jpg			

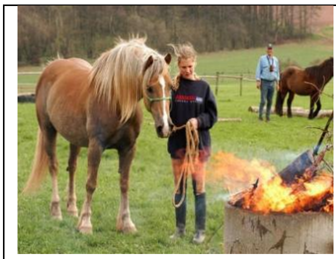
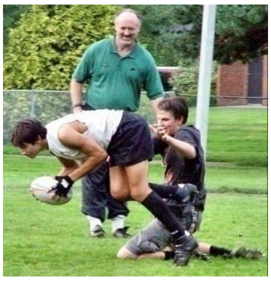
	['two people standing in the snow with skis and poles']	['two skiers are posing for a picture in the snow']	['two people standing in the snow with skis']
2960422620_81889a3764.jpg 	['a flock of birds standing on top of a sidewalk']	['a flock of birds walking around a city square']	['a flock of pigeons walking around a park']
719837187_3e7bf1d472.jpg 	["a dog running with a soccer ball in it's mouth"]	['a dog jumping in the air to catch a frisbee']	["a dog chasing a ball with a soccer ball in it's mouth"]
1245022983_fb329886dd.jpg 	['a little girl standing on a sidewalk with a frisbee']	['a young girl posing for a picture with a skateboard']	['a girl in a pink dress standing on a sidewalk']
1472882567_33dc14c8b6.jpg 	['a man riding skis down a snow covered slope']	['people on skis standing on top of a snow covered slope']	['a man riding skis down a snow covered slope']
2466495935_623b144183.jpg 	['young girls playing a game of soccer']	['a soccer game is being played on a field']	['young men playing a game of soccer']
10815824_2997e03d76.jpg	['a man standing next to a brown horse near a fire hydrant']	['two horses standing next to each other in a field']	['a man standing next to a horse in a field']

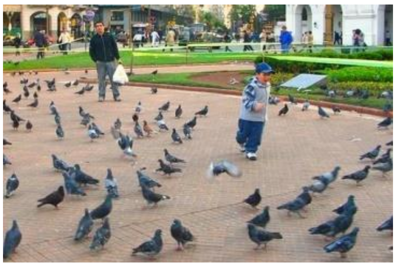
2949014128_0d96196261.jpg	['young men playing a game of soccer']	['a man kicking a soccer ball in the air']	['young men playing a game of soccer']

The above table shows different images taken from Flickr8kDataset namely 1000268201_693b08cb0e.jpg,101669240_b2d3e7f17b.jpg ,2960422620_81889a3764.jpg,719837187_3e7bf1d472.jpg,1245022983_fb329886dd.jpg,1472882567_33dc14c8b6.jpg,2466495935_623b144183.jpg,10815824_2997e03d76.jpg,2949014128_0d96196261.jpg. In order to get various descriptions of the above Flickr images, applied the mentioned methods. For example, consider the image, 2960422620_81889a3764.jpg

Image	Vision Transformer Result	K-Means Result	Mean Shift Result	Story Telling Result
1000268201_693b08cb0e.jpg	['a little girl standing next to a wooden bench']	['a pile of junk sitting on top of a wooden structure']	['a little girl standing in front of a window']	[a little girl standing next to a wooden bench,a pile of junk sitting on top of a wooden structure,a little girl standing in front of a window]
101669240_b2d3e7f17b.jpg	['two people standing in the snow with skis and poles']	['two skiers are posing for a picture in the snow']	['two people standing in the snow with skis']	[two people standing in the snow with skis and poles,two skiers are posing for a picture in the snow,two people standing in the snow with skis]
2960422620_81889a3764.jpg	['a flock of	['a flock of birds	[‘a flock of	['a flock of birds

	birds standing on top of a sidewalk']	walking around a city square']	birds flying over a crowd of people']	standing on top of a sidewalk,a flock of birds walking around a city square,a flock of birds flying over a crowd of people']
719837187_3e7bf1d472.jpg	["a dog running with a soccer ball in it's mouth"]	['a dog jumping in the air to catch a frisbee']	"a dog chasing a ball with a soccer ball in it's mouth"]	[a dog running with a soccer ball in it's mouth,a dog jumping in the air to catch a Frisbee, a dog chasing a ball with a soccer ball in it's mouth]
1245022983_fb329886dd.jpg	['a little girl standing on a sidewalk with a frisbee']	['a young girl posing for a picture with a skateboard']	['a girl in a pink dress standing on a sidewalk']	[a little girl standing on a sidewalk with a Frisbee,a young girl posing for a picture with a skateboard,a girl in a pink dress standing on a sidewalk]
1472882567_33dc14c8b6.jpg	['a man riding skis down a snow covered slope']	['people on skis standing on top of a snow covered slope']	['a man riding skis down a snow covered slope']	[a man riding skis down a snow covered slope,people on skis standing on top of a snow covered slope,a man riding skis down a snow covered slope]
2466495935_623b144183.jpg	['young girls playing a game of soccer']	['a soccer game is being played on a field']	['young men playing a game of soccer']	[young girls playing a game of soccer,a soccer game is being played on a field,'young men playing a game of soccer]
10815824_2997e03d76.jpg	['a man	['two horses	['a man	[a man standing

	standing next to a brown horse near a fire hydrant']	standing next to each other in a field']	standing next to a horse in a field']	next to a brown horse near a fire hydrant,two horses standing next to each other in a field,a man standing next to a horse in a field]
2949014128_0d96196261.jpg 	['young men playing a game of soccer']	['a man kicking a soccer ball in the air']	['young men playing a game of soccer']	[young men playing a game of soccer,a man kicking a soccer ball in the air,young men playing a game of soccer]



By directly applying Vision Transformer (ViT), getting the caption of

['a flock of birds standing on top of a sidewalk']

For the same image, applying K means technique and then applying ViT to get another caption of

['a flock of pigeons walking around a park']

Similarly, applying the Mean Shift technique and applying ViT to get one more description other than the remaining techniques getting a description of .This process is the same for the remaining images and gets different captions.

4.8 Integrate the outputs

After generating various descriptions of a same image, then perform integration of all the generated captions then it generates a story of that image.

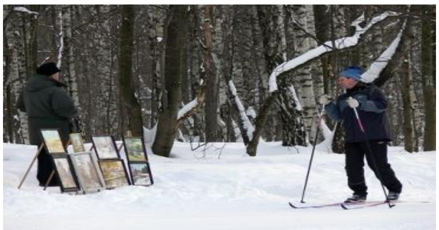
Table 6 Story Telling of images

Table 6 indicates the story telling of the images taken from the Flickr 8k dataset. Story telling is constructed by combining the all the descriptions generated from Transformer result, K-Means result

and Mean Shift result.

For example, consider one of the above images, is shown below.

101669240_b2d3e7f17b.jpg



By applying ViT, getting caption of

['two people standing in the snow with skis and poles']

After performing clustering techniques, the captions of the above image are more techniques and generate a story with coherent description.

['two skiers are posing for a picture in the snow']

['two people standing in the snow with skis']

By combining all the description, generates a story behind it. Similarly the remaining images also produce a story behind them. Then the final result

[two people standing in the snow with skis and poles,two skiers are posing for a picture in the snow,two people standing in the snow with skis]

5 Evaluation Metrics

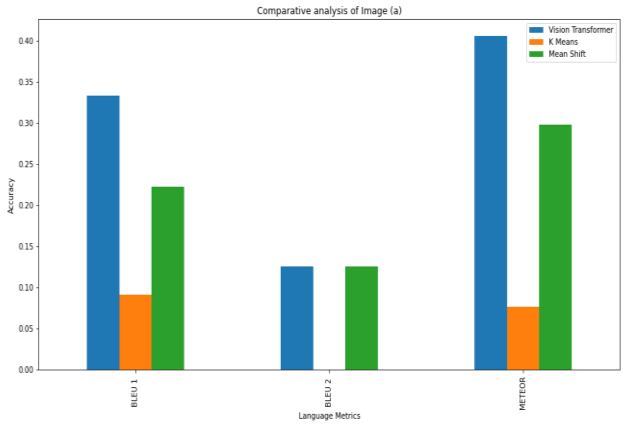
The below figures from (a) to (i) describe the analysis of the generated descriptions with language metrics of BLEU 1[Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering) [26] and other language evaluation metrics are presented to enhance the redescriptions of the same image [26]. These metrics generate the scores of the descriptions and evaluate the quality of the generated descriptions. Each score is represented in Table 7. By comparing each description with the proposed algorithms through language metrics, the following graphs are generated.

Table 7 Comparative Analysis with Language Metrics

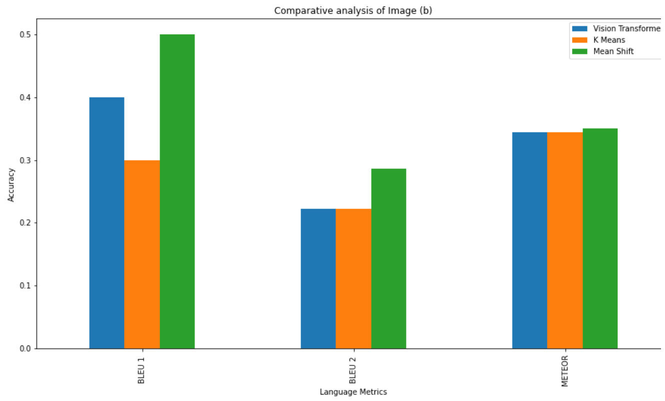
S.NO.	Image	BLUE 1	BLUE 2	BLUE 3	METEOR
1	1000268201_693b08cb0e.jpg	0.3333333333333333	0.125	0.0	0.40564373897707234
	K-Means	0.09090909090909091	0.0	0.0	0.07692307692307693
	Mean Shift	0.2222222222222222	0.125	0.0	0.29761904761904767
	101669240_b2d3e7f17b.jpg	0.4	0.2222222222222222	0.125	0.34403669724770647

2					
	K-Means	0.3	0.2222222222222222	0.125	0.34403669724770647
	Mean Shift	0.5	0.2857142857142857	0.16666666666666666	0.35046728971962615
3	2960422620_81889a3764.jpg	0.4	0.2222222222222222	0.0	0.2485236220472441
	K-Means	0.2222222222222222	0.125	0.0	0.14880952380952384
	Mean Shift	0.2	0.1111111111111111	0.0	0.07874015748031496
4	719837187_3e7bf1d472.jpg	0.3	0.0	0.0	0.0974025974025974
	K-Means	0.6	0.2222222222222222	0.125	0.33189033189033185
	Mean Shift	0.25	0.0	0.0	0.09615384615384616
5	1245022983_fb329886dd.jpg	0.1	0.0	0.0	0.28956422018348627
	K-Means	0.2	0.0	0.0	0.09174311926605505
	Mean Shift	0.2	0.0	0.0	0.28956422018348627
	1472882567_33dc14c8b6.jpg	0.0	0.0	0.0	0.2604166666666667
6	K-Means	0.09090909090909091	0.0	0.0	0.2533783783783784
7	Mean Shift	0.0	0.0	0.0	0.2604166666666667
	2466495935_623b144183.jpg	0.2857142857142857	0.0	0.0	0.3488372093023256
	K-Means	0.1111111111111111	0.0	0.0	0.22222222222222224
	Mean Shift	0.2857142857142857	0.0	0.0	0.3488372093023256
	10815824_2997e03d76.jpg	0.25	0.0	0.0	0.196078431372

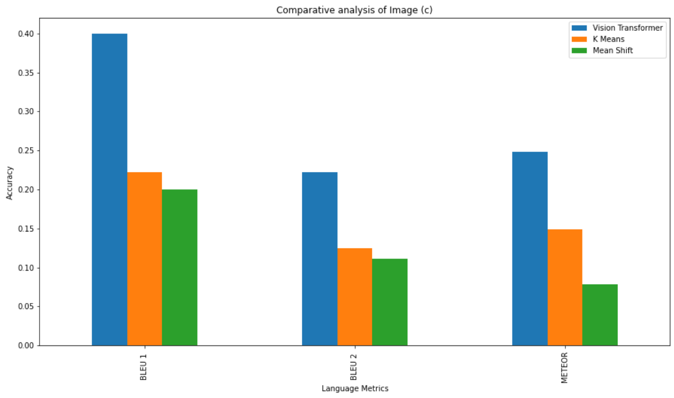
8					54904
	K-Means	0.2	0.11111111111 11111	0.0	0.187500000000 00003
	Mean Shift	0.1	0.0	0.0	0.100000000000 00002
9	2949014128_0d96196261.jpg	0.285714285714 2857	0.0	0.0	0.045871559633 027525
	K-Means	0.4444444444444 4444	0.125	0.0	0.292245370370 37035
	Mean Shift	0.285714285714 2857	0.0	0.0	0.094339622641 50944



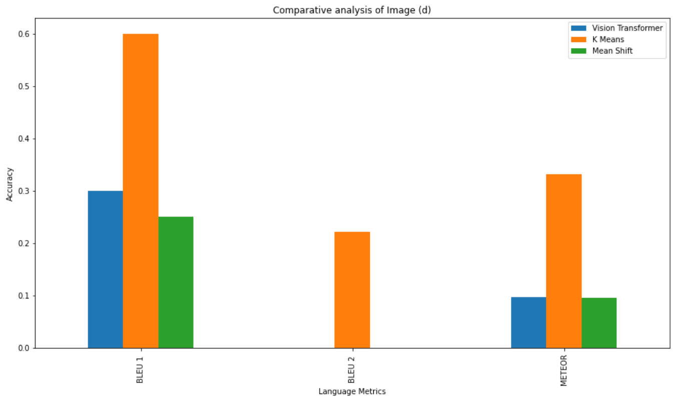
(a)



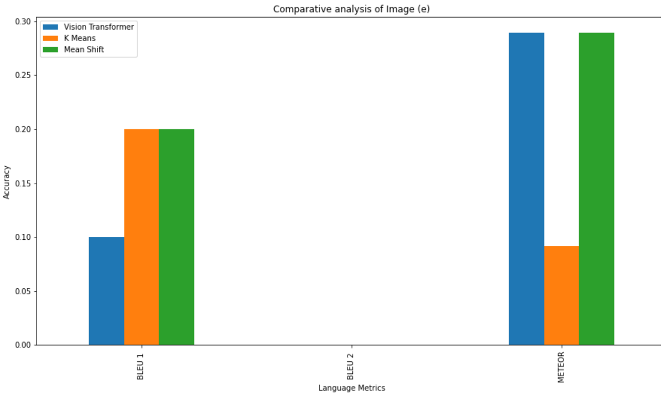
(b)



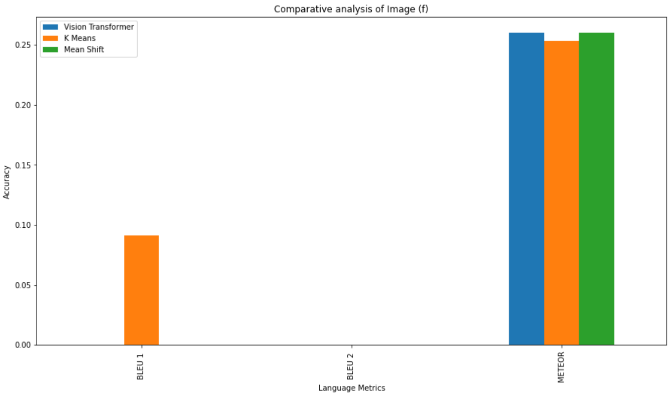
(c)



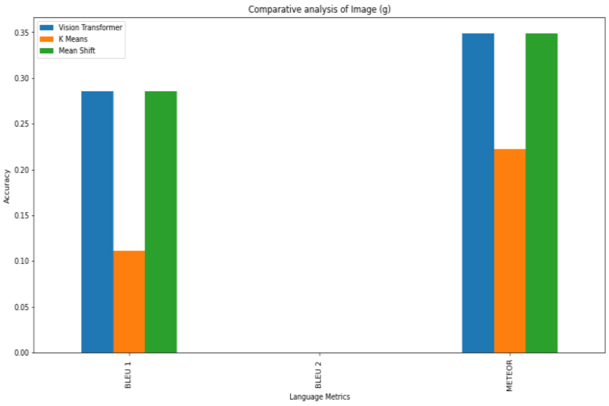
(d)



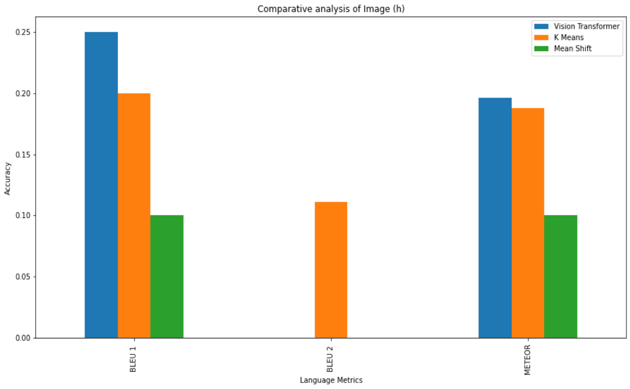
(e)



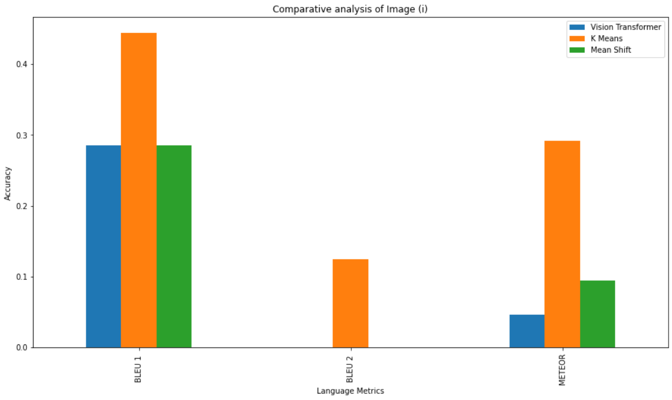
(f)



(g)



(h)



(i)

Conclusion and Future Work

This Paper constructs a story behind an image. There are some methods that describe the image in the form of text .This paper discusses about vision transformers that give the description of the image and two types of clustering algorithms. All these mentioned methods were applied in order to get various descriptions of the same image so that it generated redescription of the same image. After getting different descriptions of the same image, perform integration of all the resulting descriptions and finally conclude with a story about the particular image. In future, we can extend our approach with more techniques and generate a story with a coherent description.

References

[1] K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[2] T. Jaknamon and S. Marukatat, "ThaiTC:Thai Transformer-based Image Captioning," 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 2022, pp. 1-4, doi: 10.1109/iSAI- C. Orhei NLP56921.2022.9960246.

[3] , M. Mocofan, S. Vert and R. Vasiu, "End-to-End Computer Vision Framework," 2020 International Symposium on Electronics and Telecommunications

(ISETC), Timisoara, Romania, 2020, pp. 1-4, doi: 10.1109/ISETC50328.2020.9301078.

[4] J. Wang, Z. Chen, A. Ma and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199

[5] P. G. Shambharkar, P. Kumari, P. Yadav and R. Kumar, "Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 857-864, doi: 10.1109/ICICCS51141.2021.9432245.

[6] Y. Yang, "Image-Caption Pair Replacement Algorithm towards Semi-supervised Novel Object Captioning," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 266-273, doi: 10.1109/ICSP54964.2022.9778729.

[7] Sule Anjomshoae, Daniel Omeiza, Lili Jiang,Context-based image explanations for deep neural networks,Image and Vision Computing,Volume 116,2021,104310,ISSN 0262-8856,https://doi.org/10.1016/j.imavis.2021.104310.(https://www.sciencedirect.com/science/article/pii/S0262885621002158)

[8] J. Wang, Z. Chen, A. Ma and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199.

[9] Absalom E. Ezugwu, Abiodun M. Ikotun, Olaide O. Oyelade, Laith Abualigah, Jeffery O. Agushaka, Christopher I. Eke, Andronicus A. Akinyelu,A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,Engineering Applications of Artificial Intelligence,Volume 110,2022,104743,ISSN 0952-1976,https://doi.org/10.1016/j.engappai.2022.104743.

[10] S, D., S, Q., Y, X., S, A. & S., W. (2019). Image caption generation with high-level image features. Pattern Recognition Letters, 123:89–95. doi: 10.1016/j.patrec.2019.03.021.

[11] Ding, G., Chen, M., Zhao, S. et al. Neural Image Caption Generation with Weighted Training and Reference. Cogn Comput 11, 763–777 (2019). https://doi.org/10.1007/s12559-018-9581-x

[12] M. A. Hassan, S. Saleem, M. Z. Khan and M. U. G. Khan, "Story Based Video Retrieval using Deep Visual and Textual Information," 2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE), Islamabad, Pakistan, 2019, pp. 166-171, doi: 10.1109/C-CODE.2019.8680995.

[13] I. K. Raharjana, D. Siahaan and C. Fatichah, "User Stories and Natural Language Processing: A Systematic Literature Review," in IEEE Access, vol. 9, pp. 53811-53826, 2021, doi: 10.1109/ACCESS.2021.3070606.

[14] Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur," Image Caption Generator," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-10 Issue-3, January 2021.

[15] J. Vaishnavi and V. Narmatha, "Video Captioning based on Image Captioning as Subsidiary Content," 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2022, pp. 1-6, doi: 10.1109/ICAECT54875.2022.9807935.

[16] W. Zhang and J. Ma, "Image Caption Enhancement with GRIT, Portable ResNet and BART Context-Tuning," 2022 6th International Conference on Universal Village (UV), Boston, MA, USA, 2022, pp. 1-6, doi: 10.1109/UV56588.2022.10185494.

[17] A. Z. Al-Jamal, M. J. Bani-Amer and S. Aljawarneh, "Image Captioning Techniques: A Review," 2022 International Conference on Engineering & MIS (ICEMIS), Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ICEMIS56295.2022.9914173.

[18] V. Atliha and D. Šešok, "Comparison of VGG and ResNet used as Encoders for Image Captioning," 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2020, pp. 1-4, doi: 10.1109/eStream50540.2020.9108880.

[19] W. Kang and W. Hu, "A Survey of Image Caption Tasks," 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Nanjing, China, 2022, pp. 71-74, doi: 10.1109/CEI57409.2022.9950150.

[20] A. Ueda, W. Yang and K. Sugiura, "Switching Text-Based Image Encoders for Captioning Images With Text," in IEEE Access, vol. 11, pp. 55706-55715, 2023, doi: 10.1109/ACCESS.2023.3282444.

[21] Al Nahian, M.S., Tasrin, T., Gandhi, S., Gaines, R., Harrison, B.: A hierarchical approach for visual storytelling using image description. In: International Conference on Interactive Digital Storytelling. pp. 304–317. Springer (2019).

[22] Malakan, Zainy M., Ghulam Mubashar Hassan, and Ajmal Mian. "Vision transformer based model for describing a set of images as a story." In Australasian Joint Conference on Artificial Intelligence, pp. 15-28. Cham: Springer International Publishing, 2022.

[23] Chen, H., Huang, Y., Takamura, H., Nakayama, H.: Commonsense knowledge aware concept selection for diverse and informative visual storytelling. arXiv preprint arXiv:2102.02963 (2021).

[24] Kang, Y., Park, H., Smit, B., & Kim, J. (2022, November 17). Moftransformer: a Multi-modal Pre-training Transformer for Universal Transfer Learning in Metal-organic Frameworks. https://scite.ai/reports/10.21203/rs.3.rs-2201064/v1.

[25] Chang, Y.-H.; Chen, Y.-J.; Huang, R.-H.; Yu, Y.-T. Enhanced Image Captioning with Color Recognition Using Deep Learning Methods. *Appl. Sci.* 2022, *12*, 209. https://doi.org/10.3390/app12010209.

[26] Darapu Uma, M.Kamala Kumari, "A Comprehensive Survey and Comparison on Story Construction Techniques Using Deep Learning for Scene Recognition," International Journal of Computer Sciences and Engineering, Vol.10, Issue.12, pp.14-22, 2022.