Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies

Sravya Bhandari[1], Abhishek Kumar[2], Priya Batta[2,3,*], Shankar Shambhu[4]

[1]Liverpool John Moores University, United Kingdom
[2]Dept. of CSE, Chandigarh University, Punjab, India
[3]Amity School of Engineering and Technology, Amity University Punjab, Mohali, India
[4]Chitkara University School of Engineering & Technology, Chitkara University, Himachal Pradesh, India

*Author to whom correspondence should be addressed:
E-mail: batta.priya1@gmail.com

(Received January 27, 2025; Revised July 08, 2025; Accepted August 02, 2025)

Abstract: Image captioning is a significant area of application for artificial intelligence techniques. When a machine can interpret an image similar to humans, it indicates a higher intelligence level and comprehension of the image. This research displays advancements in real-time image collection and labeling systems using a triad of computer vision, natural language processing, and classification. The approach employs three deep learning models to generate human-level natural language descriptors, resulting in a user-friendly system. The model comprises a multimodal pipeline of deep learning architectures, enabling the extraction of probabilistic features for each object category. Our model surpasses other image captioning models, achieving a CIDEr score of 37.93% on the common MS-COCO Captioning task test baseline, thereby exhibiting superior syntactical saliency when integrated with advanced object features. Additionally, we observed that incorporating an intermediate step of clustering objects before classification enhances the final model's performance. By implementing these methodologies, we have developed a more capable and accurate model, proficient in object classification and generating informative image descriptions. Such capabilities can significantly augment human comprehension and decision-making across various applications, particularly in advancing sustainable cities and communities, fostering quality education through improved accessibility of visual content, promoting industry, innovation, and infrastructure with cutting-edge AI technologies.

Keywords: Deep Learning; K-Means Clustering; KNN classification; MS COCO; Multimodality; POS tagging; YOLO

1. Introduction

Everyone wants to get the most done possible in the time they have. Therefore, having more free time is not a predictor of success; what matters is actually getting things done. Successful people often prefer to write down their thoughts, ideas, and to-do lists, or to program their lives down to the smallest detail. In order to easily recall ideas that help them thrive and achieve tremendous success later in life, they should write them down. In a positive light, it's encouraging to realize that even routine activities, like writing down and organizing one's thoughts, can lead to breakthrough ideas. The alternative to our intervention is that they will depart on their own or become trapped in a cemetery. The five senses—sight, sound, smell, taste, and touch—are the origin of all human experience. Furthermore, a person's present ideas are the outcome of their accumulated input over time.

Data production nowadays is indeed mind-boggling, especially when compared to historical levels (about five quintillion bytes per second). Therefore, the process of studying material over and over again to store it in long-term memory is dominated by isolation. Furthermore, if we don't arrange things, they don't get revised, and if we don't get revised, we lose our ideas. The process of strategically planning and executing actions to get desired outcomes necessitates cognitive effort and determination. Suppose a someone has acquired a compilation of Ayurvedic home

Cite: S. Bhandari et al., "Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies". Evergreen, 12 (04) 1840-1866 (2025). https://doi.org/10.5109/7402620.

remedies, which are sometimes referred to as the "Mother of healing," specifically targeting the treatment of Dry Cough (referred to as Vataj Kasa in Ayurveda) or Wet Cough (referred to as Kaphaj Kasa in Ayurveda) utilizing readily available kitchen herbs as prescribed by a medical practitioner. Instincts and the surrounding environment can lead one to choose a combination of herbs like Black Pepper, Betel leaf, and Holy Basil, for instance.

When that individual next gets sick, and wants to utilize that cure, but only if the data has been properly categorized in a selectable manner to ensure a good match between the symptoms and the medicine. One must be able to pick a useful nugget of information out of a plethora of data when the time comes. Therefore, it is crucial to organize the data with the help of the available tools and technologies, no matter how large the data is. Machine learning has played a substantial role in facilitating the transformation of abstract concepts into practical applications, particularly in the context of organizing diverse collections of information. It has proven effective in recommending suitable labels for categorizing unwieldy sources of data[1)]. This could be done for a variety of reasons, including the dissemination of knowledge or the actualization of ideas. Therefore, it is crucial for efficient cross-modal retrieval to establish optimum multi-modal data representations[2)]. Instead of representing image-text pairs as unified feature vectors in a shared representation space, as is done in conventional methods for image-text retrieval, our solution incorporates generative processes into the cross-modal feature embedding[1-5)]. Using this method, we are able to learn both global abstract picture features and local grounded features at the same time, which leads to improved cross-multi-modal retrieval on the MS COCO dataset.

1.1. Motivation

The belief that data is crucial to knowledge acquisition and informed decision-making drives this research. Large amounts of data do not guarantee their usefulness or in formativeness. In this circumstance, an excess of data can exist without equivalent information, emphasizing the difference. The contextual framework emphasizes the significance of decluttering and purifying data, as unstructured and unprocessed erroneous data is useless. The massive amounts of data being generated must be efficiently organized and used to gain insights and make informed decisions. This is crucial for time- and resource-conscious individuals and businesses. Machine learning and technology will be used to streamline this process by providing tools and methods for organizing, classifying, and retrieving data. The paper also proposes incorporating generative processes into cross-modal feature embedding to improve data retrieval. This method improves global abstract picture characteristics and local grounded image features for more accurate and efficient data retrieval. Thus, people and corporations may better manage their time and resources.

1.2. Contribution

This study's multi-modal pipeline was accomplished by fusing computer vision and Natural Language Processing techniques to integrate cross-modality properties of images and texts; the study's rich object features aided in the generation of more precise captions. A further proposal used clustering followed by classification to quickly improve accuracy from 75.43 to 82.54% by separating the correlated categories from the captions of the photos created.

1.3. Organization

Here is how the rest of the paper is structured: Section 2 delves into the literature, Section 3 outlines the research approach, Section 4 provides experiment details, Section 5 discusses implementation, and Section 6 elaborates on the findings. The concluding section will focus on future perspectives.

2. Literature Review

Using the captioning framework, the authors describe a real-time captioning engine that generates natural language descriptions on par with a human's. The development of this system necessitated the use of multimodal translation[6,7)]. The final system should be user-friendly and automatically annotate massive MS COCO collections[4)]. Our method also extracts characteristics from prominent class items in an image to build picture captions with data descriptions available via computer vision models like YOLO[8,9)]. Natural language models based on probabilistic feature weights would designate image object-derived word objects as nouns in a folder of categorized photos[10,11)]. While most recent research incorporate this multimodality with a multi-label classification of class categories, image captioning modeling is challenging since it necessitates knowledge of both Computer Vision and Natural language modality[12-14)]. The ultimate goal of image captioning is to automatically construct natural-sounding tales. We can harmonize captions and images with this. We then train a weakly supervised object detection system. Transfer learning methods should eliminate the heterogeneity gap between modalities and allow the final generative language to naturally and grammatically communicate image objects and connections[15-18)]. To create the source sequence's "visual words" from conspicuous features, the items are arranged sequentially and mapped onto a frequently hidden region[19)].

In summary, this is a multi-modal translation problem with multi-label classification on MS COCO dataset classes[20,21)]. The recognized picture components' words are translated into the target language in the source input phrases. Multi-modal data makes it hard for users to find

Cite: S. Bhandari et al., "Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies". Evergreen, 12 (04) 1840-1866 (2025). https://doi.org/10.5109/7402620.

and use heterogeneous information efficiently. This study addresses the multimodal information retrieval problem of retrieving sentences that label visuals. The most significant challenge to cross-modal retrieval is the existence of heterogeneity in data modalities.

3. Methodology

This chapter uses the coupled image and text data of captions from the MS-COCO dataset to demonstrate how object identification and picture-captioning models are structured. Manually collecting enough labeled training data is laborious, and retraining deep captioning algorithms is a time-consuming process. Therefore, we analyze pertinent written and visual sources to find potential solutions to this problem[22-24]. By reviewing the pertinent literature, we hope to determine which methods have proven most useful for our model in past studies. We build a multi-label classifier-based object detection module, a captioning model, and a captioning model module in this research. To determine polarity in images, a label classifier is used in conjunction with an object detector to compile the various pairs associated with objects, while a captioning model is concerned with essential descriptions that identify significant visual components[25-28]. Then, we create interesting descriptions by blending factual details with emotional conceptions of probability commonalities.

3.1. Dataset

To showcase our method, we use the publicly available MS COCO dataset.The COCO dataset is used for its large, diverse collection of real-world images with rich object annotations and multiple captions, making it ideal for training and evaluating image captioning models. 80,000 photos, 5 full captions, and 80 item categories make up the train, validation, and test datasets for cross-modal retrieval, which are split 8:1:1. This study makes two significant contributions. The connection between the two modalities is often observed in these models through the use of concrete, rooted representations. In addition, the empirical results of our extensive experiments on the benchmark data show that integrating grounded and abstract representations improves the performance of SOTA's cross-modal image-caption retrieval. The below Table 1 dissects the evolution of research over time, highlighting challenges addressed in past studies and the overarching purposes that guided previous investigations.

Table 1: Year wise Citation-Problem

Year	Citation-Problem	Purpose
2022	To address the multiplicity of, a single-turn MS-COCO machine-in-the-loop (MITL) annotation pipeline was implemented[16].	Use a single-turn MITL annotation process to solve MS-COCO annotation problems.
2022	Fuse picture target regions and labels to quickly complete image features and label co-occurrence embeddings[17]	Integration of picture target regions and labels through efficient fusion improves image feature completeness.
2022	Vanishing gradient[18]	Addressing image processing gradient disappearance and fine-grained captions.
2022	A low convergence efficiency[19]	Improve multi-label picture performance by increasing convergence efficiency.
2022	With high-performance scaling and more training data, text tokens and Masked Auto Encoders align picture patches for Vision-Language Captions [20]	With more training data, text tokens and Masked Auto Encoders can create "Vision-Language Captions" for flawless image patch alignment and scalable performance.
2022	To propose the fusion network model [21]	Introducing an image captioning fusion network model that outperforms others.
2022	Pixel-input image-to-sequence issue [22]	To solve the image-to-sequence challenge, encode pixel inputs into feature vectors and generate sequences using a vocabulary.
2021	New multi-label BG Capsule proposition[23]	Introducing the innovative multi-label BG Capsule for state-of-the-art text categorization without external datasets.
2022	Classification improvement[24]	Enhancing classification performance on long-tailed datasets with LSE-Sign loss function.
2020	Reduce the problem of not recognizing useful photos[25]	Addressing picture recognition limitations and their potential applications.
2019	Improving image captions beyond picture classification, object detection, and context detection[26].	Enhancing picture captions beyond image classification, object detection, and context detection.
2020	Multi-modal machine translation and language understanding[27]	Introducing a novel method for multi-modal image captioning using machine translation and language

		understanding.
2020	To build relationship Learning relationship-aware visual representations for picture description interest-areas while considering historical context and prior attention[28].	Establishing relation-aware visual representations for picture focus areas, including historical context and prior attention.
2019	Different detectors' objects are misaligned[29].	A Multi-View Image Encoder model aligns and corrects detector misalignment.
2019	A novel text classification method that treats input text as a picture[30]	A novel text classification method that leverages input text as an image to derive semantically significant characteristics without OCR.
2018	Multi-modal information retrieval embeds generative processes within cross-modal feature embedding[31].	Discussing cross-modal feature embedding for multi-modal information retrieval with generative processes.

3.2. Traditional Image captioning System

The models' outputs are referred to as a concatenation phase due to the attention-based encoder-decoder design of our model, which produces a feature matrix. This matrix enhances the language decoder's ability to accurately anticipate descriptions by providing more information. In contrast to prior work, we directly use object layout data rather than inserting object characteristics and then merging them with CNN features. Therefore, we use a Bahdanau attention module for language generation that consists of a GRU and two fully linked layers[29-31].The Bahdanau method introduces such an attention mechanism, as a result of which the model learns to concentrate on the most significant sections of an image during the production of every word in a caption. It dynamically gives pricing weights to various regions instead of treating all features of an image equally, thus enhancing contextual precision, and semantic accuracy. It prompts captions that are more descriptive and human-like in nature and thus a significant part in the use of image captioning models.

In this study, we argue that object properties like class, size, and position can improve viewers visual imagery perception. Using picture segments, attention weights were assigned to pixels matching to each generated word to create the caption example. Figure 1 shows the highest attention weight region identified by this approach[32]. The data could improve a computational model of two frisbee players.

A generated caption utilizing specific image segments, accompanied by visualizing attention weights assigned to individual pixels for each generated word, highlights the region with the highest attention weight[33].

3.3. Image Encoding

This research makes use of the Xception CNN, the most recent iteration of InceptionV3, to extract spatial information through 71 layers of modified depth-wise separable convolution. Xception CNN is essential when encoding each image to extract deep spatial feature by utilizing depth wise separable convolutions. We used the Xception technique to isolate unique features from real-world observations, which involves sequential data processing through entry, middle (repeated eight times), and exit flows, as illustrated in Figure 2 and Figure 3.

Recent image captioning and CNN model research has examined applications that use attributes from the layer before the fully linked layer[14,34]. The model can now learn about image objects and connections, not simply the picture class.

3.4. Language Decoding

Most current studies use an encoding-decoding method based on a more intricate CNN or RNN network structure. CNNs are important since they are used to extract deep visual features of the input images, detail of objects, space areas and textures. The basis of caption generation is these encoded features. CNNs such as Xception can be used to improve the accuracy of the model resulting in the recognition of a greater number of objects and contextual descriptions of much more information in the image.

At each time step, the input image vector and the predetermined phrases are used to construct a description, for which the output word is used. Since decoding a GRU takes very little time and storage space, we use it[35,36]. Words for each frame of a caption are generated using a combination of a context vector, the frame's prior hidden state, and previously generated words. The model is trained in a predetermined fashion via propagation.

3.4.1. Bahdanau Attention

Bahdanau's design document This deterministic method of paying attention allows for continuous, non-linear processing. The term "attention" is used to describe the practice of using a method that mimics the effects of focused mental activity. This effect emphasizes key features while downplaying less crucial ones in a given image. This theory proposes that the network prioritize the processing of a small but critical portion of the available data.

Using a dataset's training data and gradient descent, the most important piece of information can be isolated.

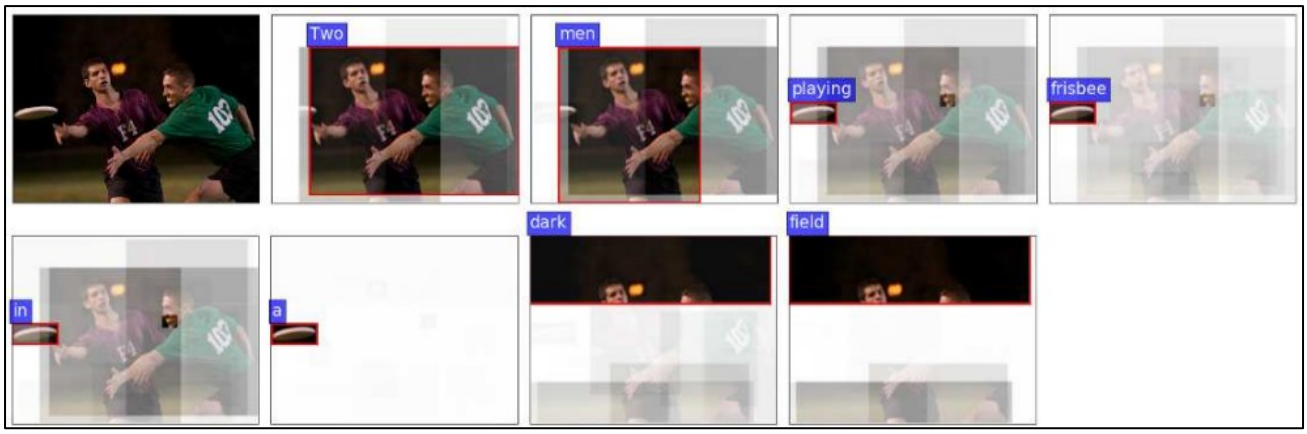

Fig. 1: Importance placed on a picture of two guys throwing a Frisbee around

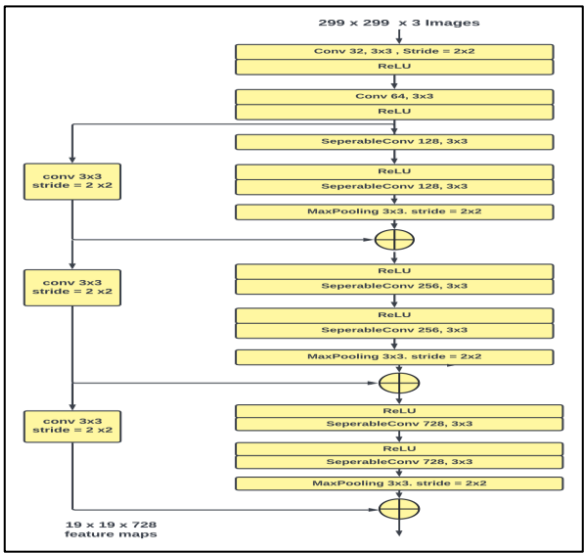

Fig. 2: The entrance flow of the Xception architecture[14)]

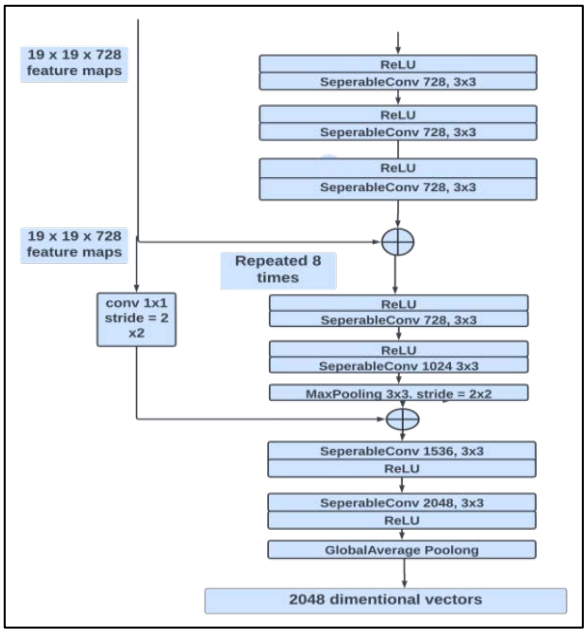

Fig. 3: The middle and exit flow of the Xception architecture[14)]

Several machine learning tasks in the fields of natural language processing (NLP) and computer vision (CV) call for human-level attentiveness. The picture captioning model is illustrated with examples from the MS COCO dataset in Figure 4, Figure 5 and Figure 6.

Our machine translation architecture benefits from this method because of how well it performs. To give the decoder more leeway in using the most pertinent parts of the input sequence, the attention mechanism integrates all encoded input vectors into a weighted combination and assigns large weights to the most relevant vectors.

Object identification features in Figure 7 and Figure 8 are indicative of the idea that knowing item classes and placements allows for a deeper understanding of a picture than simple convolutional features. When both types of information are taken into account, the algorithm will emphasize distinct characteristics of different kinds of objects and different spots within the same image.

3.4.2. YOLO Bounding-Box Detection of Objects

Incorporating additional visual information utilizing the object information present in the image is a smart way to balance the source and target sequences of recurrent neural networks. In this research, we hypothesize that the identified elements' high-level features improve the visual component of the source sequence. Sorting object characteristics by saliency increases the number of time steps in the image RNN encoding phase[37)]. The recovered features include the object's X, Y, width, and height. To determine if the box includes an item, the confidence rate and border box accuracy must be checked. YOLO-generated object attributes' object dimensions and confidence scores were fully utilized as depicted in Figure 9[38)]. Each cell's offsets are the x and y grid coordinates of its bounding box, so x, y, w, and h are all in the range 0–1. To test how well an object identification technique performs, we compare its predicted bounding box to the real world's ground truth using IoU as shown in Figure 10

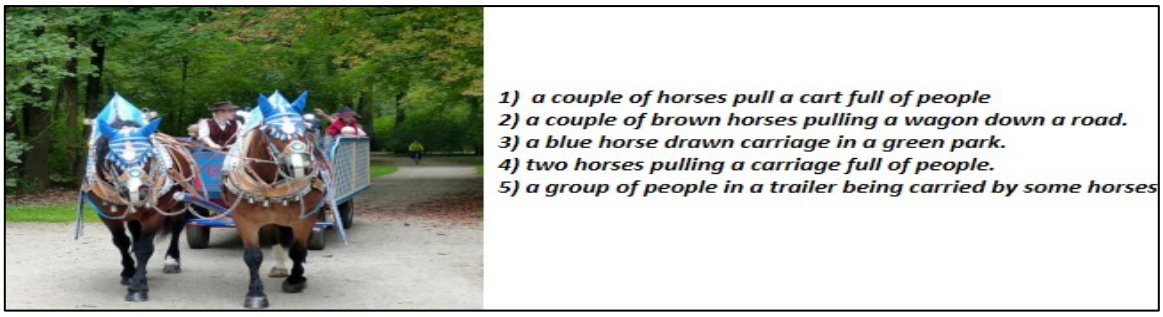

Fig. 4: Human, equine, and bicycle Classifications of images and captions using MS-COCO (using the MS COCO dataset Explore as a reference)

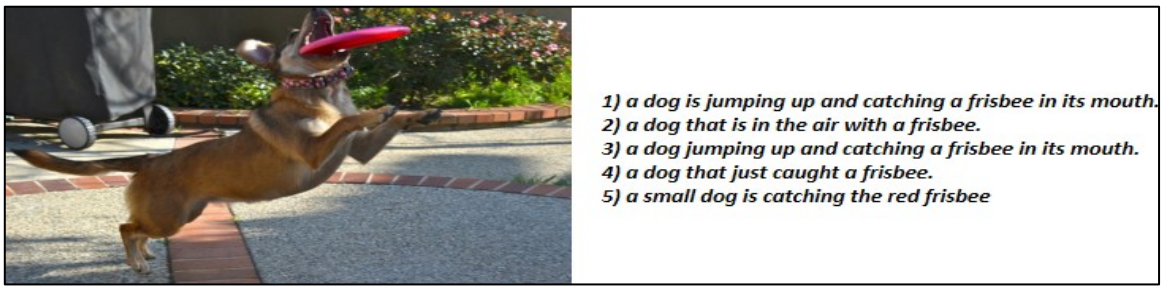

Fig. 5: Images of dogs with frisbees annotated in MS-COCO (from the MS COCO dataset Explore)

Fig. 6: Images and descriptions of furniture and home decor from MS-COCO's person, plant, vase, sofa, and chair classes

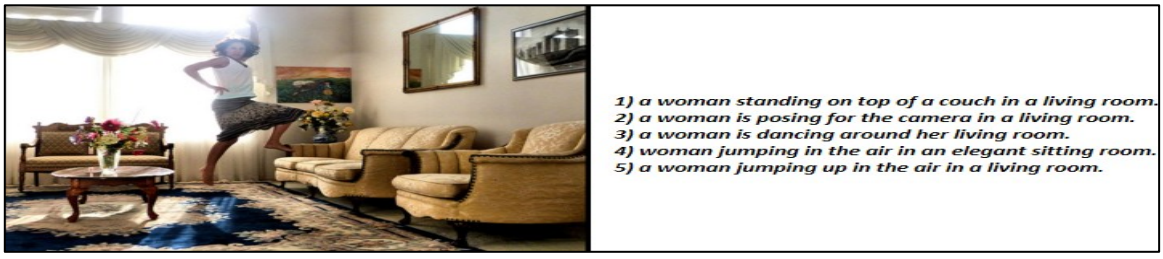

Fig. 7: Tensorflow Images captioned picture of a surfer riding a wave

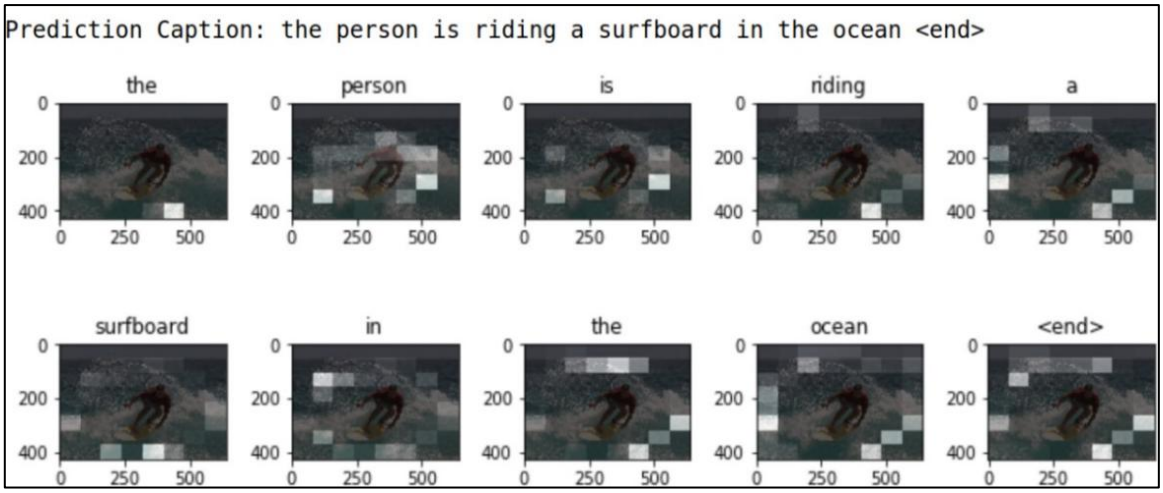

Fig. 8: This attention-based model uses TensorFlow Images along with picture regions where it focus

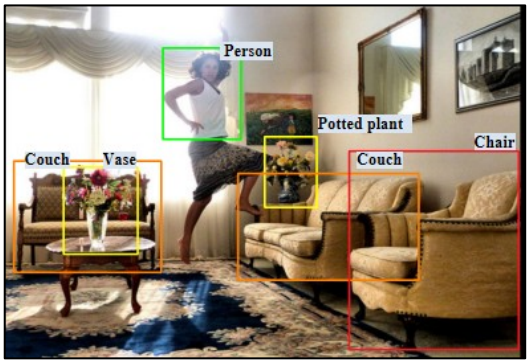

Fig. 9: Raw objects like person, potted plant, vase, couch, and chair detected from the MS-COCO image

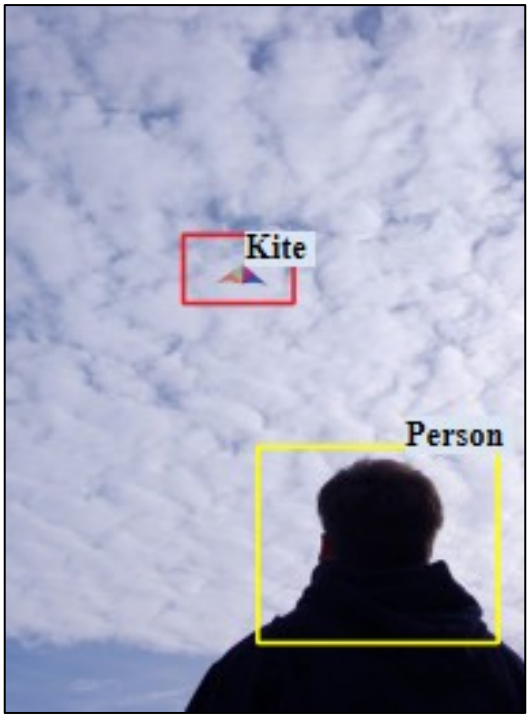

Fig. 10: Captioning the featured classes in the MS-COCO dataset, such as person and kite (retrieved from exploring the MS COCO dataset)

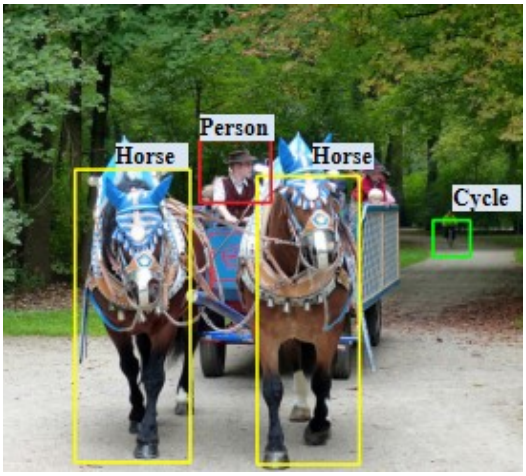

Fig. 11: Captioning the featured classes in the MS-COCO dataset, such as person

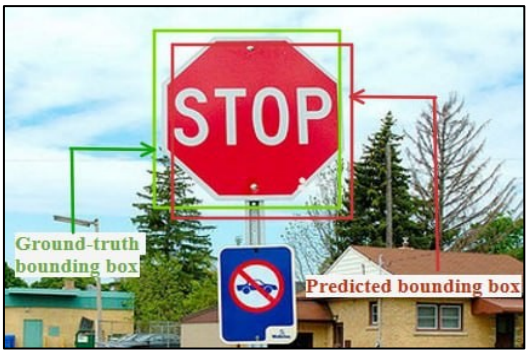

Fig. 12: The anticipated and actual bounding boxes with Intersection over Union (IOU)

and Figure 11. The method then calculates the degree of agreement between the predicted and observed boxes. The Significance score is calculated using equation 1:

$$Confidence\ score = Pr(Object) * IoU \quad (1)$$

The confidence score for each bounding box is the test result produced by the neural network. Instead of

recalculating every cell, it is used to zero in on the ones where the user has the most faith in the result as shown in Figure 12.

$$Classi(Object)\ Pr(Classi)\ Pr(Classi)\ Pr\ (Classi)\ = IoUtruth \quad (2)$$

Each prediction's confidence interval depends on width, height, x, and y. The grid cell's geographical center is (x, y). Width and height are calculated using the complete picture. The unpaid difference between actual and anticipated portions is the confidence forecast. For each grid cell, C conditional class probabilities Pr(Classi Object) are predicted using equation 2.
Where anything falls on a grid determines its likelihood. We estimate a single class probability per square independent of B. Box confidence estimates are compounded by conditional class probabilities at testing.

3.4.3. Proposed Methodology – Pseudo reference

Table 2: Pseudocode for object feature model

```
# Step 1: Extract features from
images
- Get unique images and filter
unprocessed ones.
- If there are unprocessed images:
 - Create batches of images.
 - Extract features.
 - Perform POS tagging.
 - Clusterize and classify.
 - Combine YOLO features with
clustered features and save.
# Step 2: Evaluate model
- Load encoder and decoder paths.
- Load COCO dataset.
- Create COCO evaluation object.
- Evaluate results and print scores.
# Step 3: Image Captioning Model
- Initialize model attributes.
# Step 4: Evaluate Image
- Initialize variables.
- Process an image:
 - Extract YOLO features.
 - Combine image and YOLO features.
 - Generate captions using the
model.
# Step 5: Perform POS Tagging and
Feature Extraction
- Given features and batch
information:
 - Perform POS tagging on captions.
 - Extract Noun tags from object
detection labels.
 - Combine interpretations from POS
tagging and Noun tags.
# Step 6: Clusterization and Multi-
label Classification
- Given POS tags and features:
 - Cluster different phases.
 - Perform multi-label
classification Categories-Sub
categories.
# Step 7: Main Process
- Extract and save features.
- Evaluate the model.
- Initialize Image Captioning Model.
- Evaluate images and generate
labels, then print the results.
```

4. Pre-processing and Embedding

Generate structures that link words to indexes and indexes to words. Using them, token sequences are converted into word identifier sequences.
The need for cushioning is necessitated by the variable length of sentences, which necessitates inputs of the same size. In addition, null tokens are added to the end of identifier sequences to ensure that their lengths are uniform.
To make use of the model's picture classification and object identification features (created with Lucid), output is illustrated in Table 2 and Figure 13.

5. Experimental Implementation

The data analysis, main visualizations, and procedural procedures used to develop the final approach as shown in Figure 14. For picture captioning and YOLOv4, we use an embedded model-based word extraction process. Data analysis was done with NumPy and Pandas, and visualization with Seaborn and Matplotlib. The models were built using scikit-learn, TensorFlow, Keras, YOLOV4, Scikit-learn, and NLTK.

5.1. Dataset Exploration and Core Statistics

- Microsoft Corporation's MS-COCO is a huge dataset for object detection, segmentation, and captioning. Common Objects in Context, or COCO, is an abbreviation.
- Images in the collection were captured in their natural contexts and feature daily items. The 1.5 billion object instances in the dataset are broken down into 80 different categories to guarantee precise object localization. The datasets included are of the highest quality, and they mostly make use of state-of-the-art neural networks.
- bbox: A COCO bounding box specified by its width, height, as well as its upper left x and y

Cite: S. Bhandari et al., "Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies". Evergreen, 12 (04) 1840-1866 (2025). https://doi.org/10.5109/7402620.

coordinates.

- In the COCO Bounding box, the x-top left, y-top left, breadth, and height values are saved.
- iscrowd: iscrowd=0 is used for single-object segmentation, and iscrowd=1 is used for RLE if the image contains several items. RLE is a compression method in which the frequency of occurrence is substituted for a repeated value.
- When a single item (iscrowd=0) is blocked by another, it may take more than one polygon to reveal it.
- For instance, 0 11 00 1111 0 becomes 1 2 2 4 1. Each of the category ids is unique. If we have enough information to classify food as a supercategory, then we may say that a category falls under that class. If that's the case, we can classify broccoli, doughnuts, and sandwiches as different types of food.
- Annotation format is used to hold image captions. Each picture has at least five descriptions.

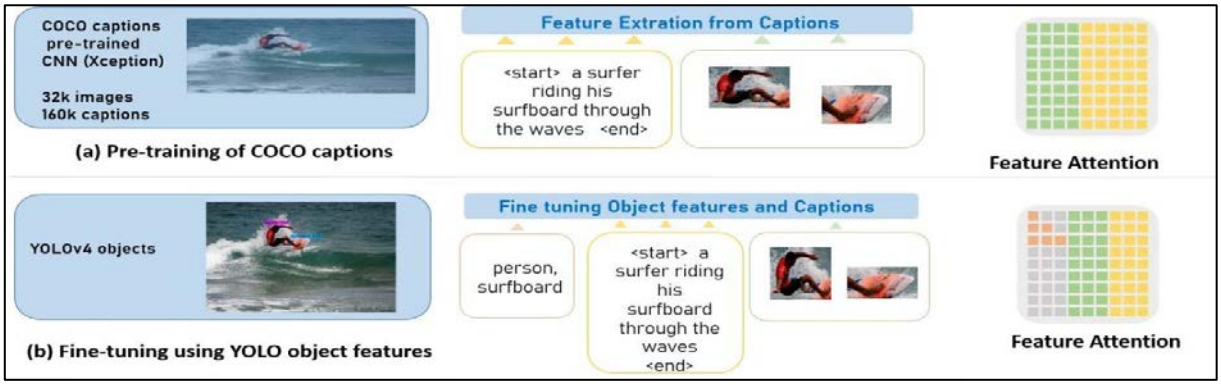

Fig. 13: Embedded caption and object feature model illustration YOLO

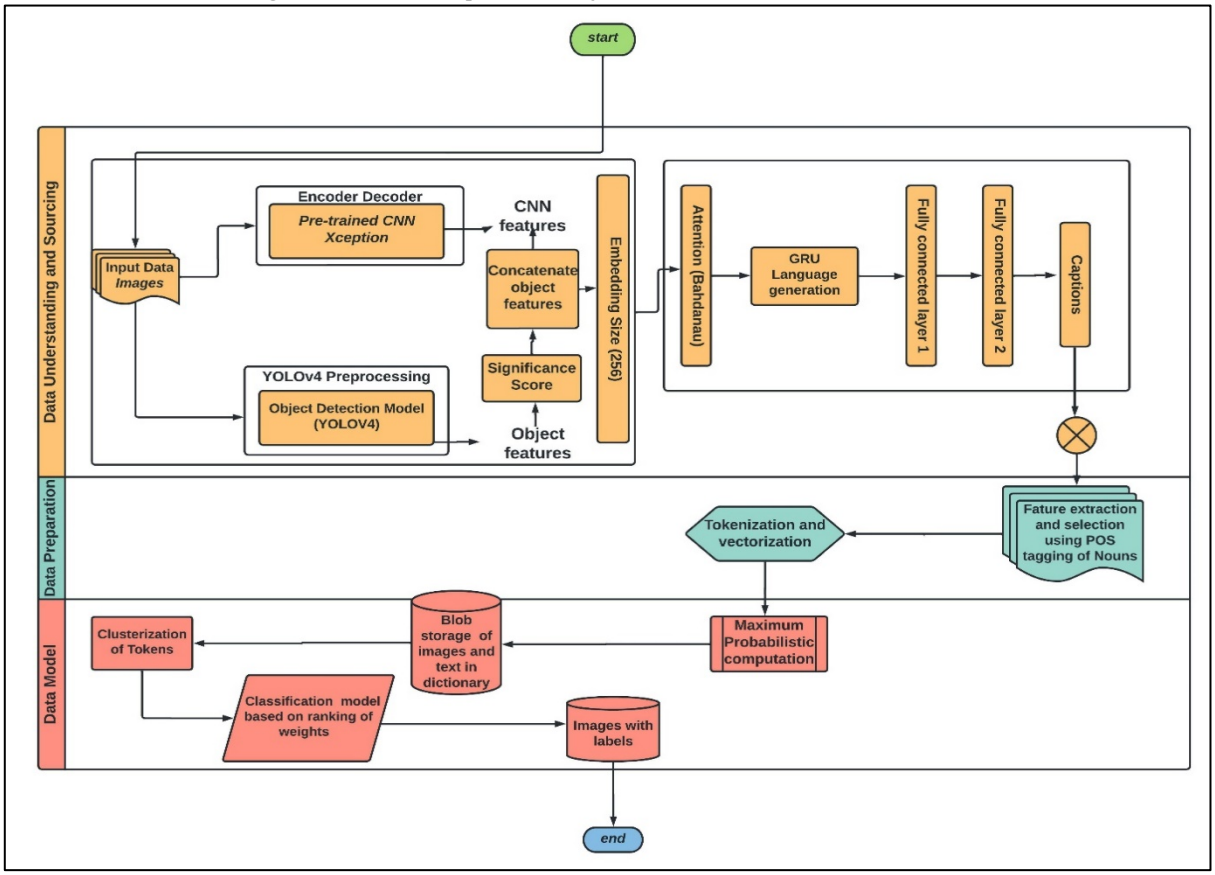

Fig. 14: Overall pipeline flow diagram of our methodology

Table 3: Data split

Dataset	Total Count of Images	Training Data Split	Validation Data Split	Testing Data Split
MS COCO	40000	32000	4000	4000

5.2. Data Selection and Fundamental Data Exploration

- The utilization of infographics as a means of data visualization facilitates the comprehension and interpretation of data by presenting it in a visual format that provides contextual information. The comprehensibility of data by the human mind facilitates the identification of trends and patterns within extensive datasets.
- We can quickly and readily understand the data because it has been organized in a way that is straightforward, and the tendencies that emerge from this organization have been highlighted.
- In this study, our model underwent training with a dataset comprising 40,000 images randomly selected from the MS COCO dataset. This set included five distinct captions and covered 80 different class categories. Table 3 outlines the 8:1:1 split for training, validation, and testing, demonstrating the balanced contribution of each class category to the overall cross-modal retrieval. It is important to note that, in consideration of unique implementation methodologies and time constraints, we constrained our use of the MS COCO dataset to a subset.

Figure 15 uses a word cloud to demonstrate how the selected dataset spans a variety of class dimensions. Each word's magnitude here reflects its relative frequency and importance in this context. This statistic reveals how often certain types of outcomes occur within a specific set of captions.

To ensure a comprehensive understanding of the underlying data distribution before applying machine-learning models, we employ a distribution graph. This graph not only indicates the frequency of potential values but also illustrates the overall distribution of labels across the chosen train-validation-test split. The dataset in our research comprises all 80 class categories, as illustrated in Figure 16.

5.3. Overview on Automated Image Labeling using Attention-Based Multimodal

More than one modality may be required to address or deal with the vast majority of the problems we face in the real world. It could include a wide variety of modality combinations. Systems that rely on numerous modalities in order to make a consolidated conclusion are inherently more resilient, efficient, and dependable than those that rely on only one.

Not only the visible component of a picture taken into account when an image recognition system is trying to identify a specific notion. On occasion, though, we also factor in things like aura, smell, and tactile experience. As a result, we rely heavily on other signals we can pick up in addition to our eyesight. Humans excel over particular AI models in this situation because of their superior capacity to distinguish specific individuals. Humans use a wide range of senses and perceptions, including smell, touch, hearing, voice, and sight, whereas AI models tend to rely on only one.

Utilizing the COCO dataset, our research develops a multi-modality model using image annotations and object recognition. Our model employs an attention-based encoder-decoder framework, combining YOLO object detection in image captioning (Encoder), a GRU-based Decoder, POS tagging, and K-means clusterization for label classification/segregation.

Figure 17 exemplifies the final model pipeline, illustrating various phases of the overall architecture designed in Lucidchart.

Feature extraction targets are combined with the language decoder model's feature matrix for predicting appropriate

Fig. 15: An illustration depicting a word cloud of the COCO dataset

descriptions. In the multimodal system, critical features, such as YOLO object detection and image captioning, are represented through early fusion[39-42].

5.4. Model Implementation of Novel Pipeline

When could it be possible for a machine to "understand" an image? The model's ability to create labels that highlight crucial elements in an image including potentially separate items is one possible interpretation. Awareness of the image's contents is required for identifying the salient material, while a practical comprehension of the scene's thrilling or pertinent aspects is required for identifying the image's potential. In this experiment, a new approach is used to generate captions for images.

We used a deep convolutional neural network (CNN) with a YOLOv4 model to produce accurate predictions across a wide range of classes and labels[43]. In addition to using categorical cross-entropy and GRU labeling, we developed our own categorization schemes.

For MS-COCO photos, the GRU label captioning is taught to recognize specific visual characteristics. A CNN encoder and GRU decoder were used to create an architecture. The CNN encoder and GRU decoder were trained using ImageNet data to fine-tune the architecture[29].

The primary concerns of this research are:

- Using Encoder-Decoder Architecture to build upon the MS-COCO dataset;
- With the VIVO (pretrained ImageNet (Xception V3) model) technique, we preprocess and cache a subset of images.
- At the same time, it uses YOLOv4 processing weights to construct objects based on fundamental object attributes including object class confidence and size, as recognized from photos.
- The subsequent stage in the natural language processing pipeline, following picture pre-processing and preceding YOLOv4, is the part-of-speech (POS) tagging of proper nouns.
- Clustering picture and label similarity and classifying class categories and super categories are the study's main focus.
- The final assignment necessitates the use of the Bahdanau attention-based model, letting us see which parts of the image are most important to the model when it comes to assigning labels.

5.5. Our model architecture has been revamped based on[44,45]. Utilize the Xception CNN encoder to pre-process the images

Data pre-processing is crucial to machine learning since it unifies the data and makes it easier for the system to generate. To restate, specific algorithms can rapidly examine and understand selected data features (such as phrases, traits, and characteristics). The objective variables or expected outputs in this case are represented by the captions and are what the model is trained to predict.

In the first step, we transform Xception into the desired format by:

The image format utilized for training Xception is in accordance with the preprocess input method's normalization, which involves scaling the image's pixel values to fall within the range of -1 to 1.

Initializing Xception loads pre-trained Imagenet weights.

The Keras implementation of TensorFlow uses the final convolutional layer of the Xception architecture as the output layer, which is 10x10x2048. Over the network, images are transformed to dictionary-style vectors (image name -> feature vector) and saved. Our architecture uses Xception, a version of Inception that replaces modules with depth-wise separable convolutions.

Like the film Inception, neural networks utilize skip connections in addition to several convolutional and max-pooling blocks inside each layer. The progress of convolutional neural networks (CNNs) has been driven by the integration of layers designed for spatial subsampling and feature aggregation.

The Inception model is able to learn complicated representations with fewer parameters thanks to its modular architecture and many feature extractors. On a wider dataset, Xception outperforms Inceptionv3 by a significant margin when it comes to image classification. The model parameters are the same as in Inception, but the computing efficiency is better as shown in Figure 18.

Separable convolution layers beat conventional ones in memory utilization and processing time. Regular convolution alters the image significantly. The depth-wise convolution of separable convolution affects the image once. Computational resources are saved while evaluating updated images since they are not repeatedly altered.

In interpreting the visual attention mechanism, the Xception model. How the network of object detection encoders can generate candidate alignment, targets and How well can the model focus on the image's most prominent features?

It also allows qualitative analysis of caption results, especially for cases where the expected output does not match the desired captions. The pre-trained ImageNet weights are based on the image and initial textual input or captions. The final model takes these two types of information and assigns weights to various pixel values. After an adapt iterator splits the caption descriptions into words and develops a vocabulary of the most often used 5,000 phrases, the Text Vectorization layer takes over and turns the text into numerical sequences. Words are size-encoded lists. The offered "word-index" and "index-word" dictionaries are said to contain every term from the created vocabulary dictionary [46-48]. For each caption, words are

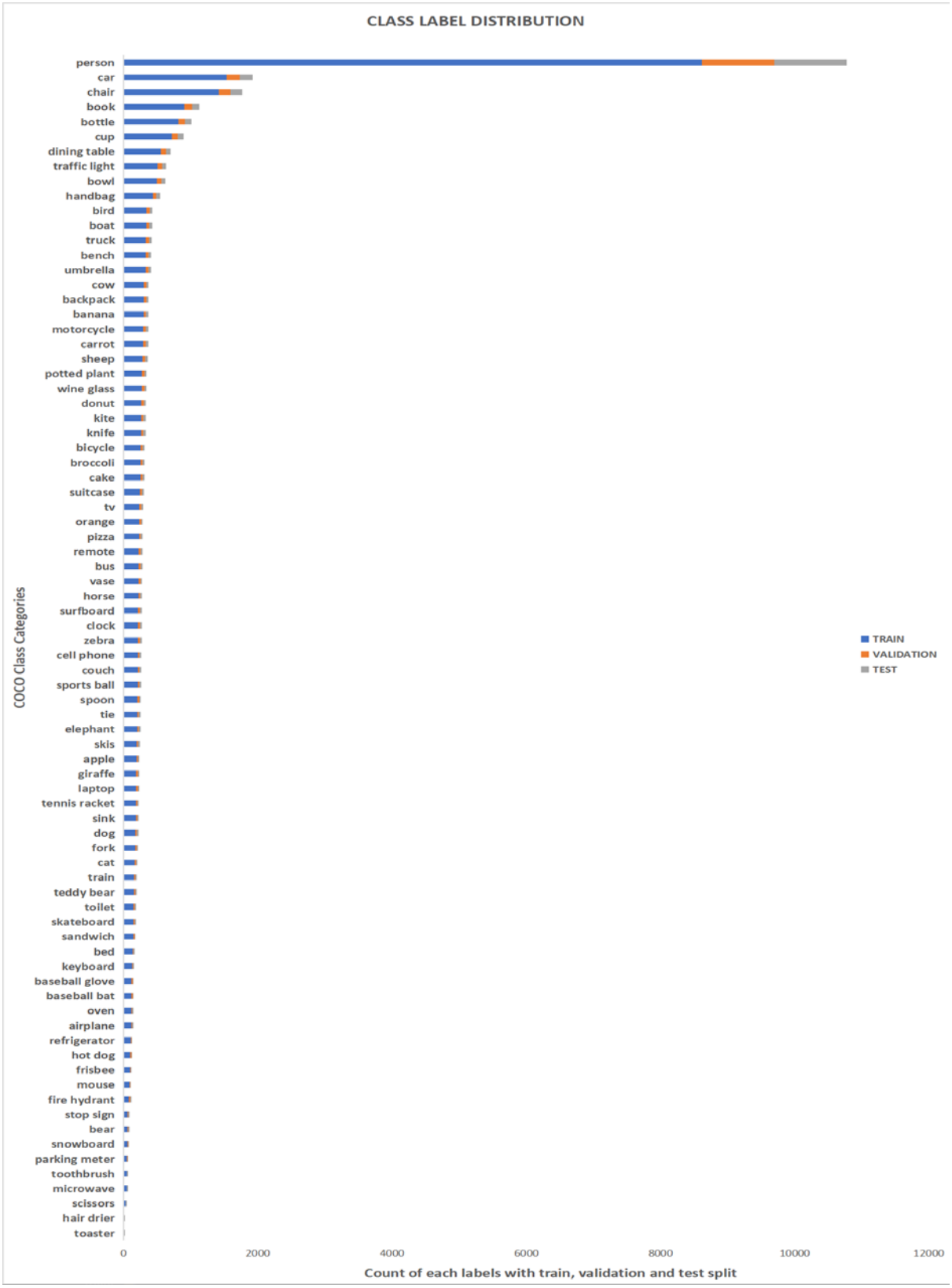

Fig. 16: Class label distribution as it appears in our model, illustrated

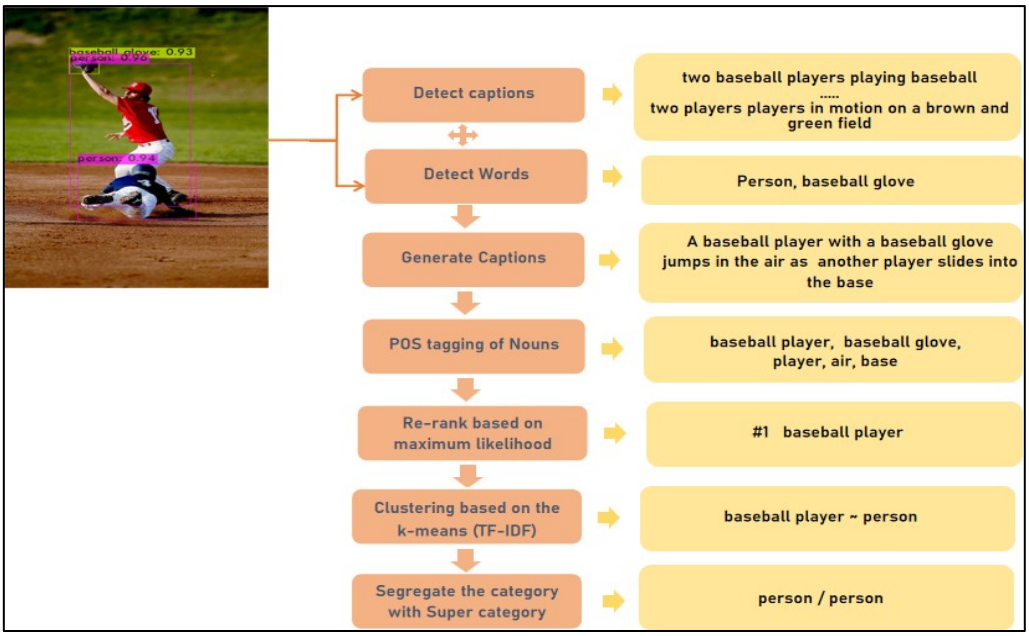

Fig. 17: Modeling workflow depiction

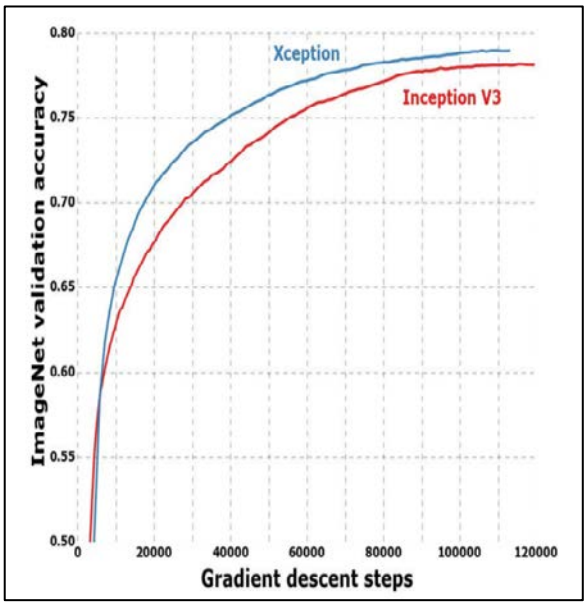

Fig. 18: Analysis of Inception and Xception's Performance on ImageNet[14)]

tokenized by assigning a numerical index based on their lexical position. Padding all output sequences to 50 generates word-to-index and index-to-word mappings. The next output word can be predicted from each output word [49,50)].

5.6. YOLOv4 Object Detection model

Most state-of-the-art models require training with a large mini-batch size, which in turn requires the utilization of many GPUs. To solve this problem, YOLO v4 created a detector that only needs a single GPU and a small minibatch size to be trained. Based on the results of a study that compared YOLOv3, YOLOv4, and YOLOv5 in terms of performance, it was determined that YOLOv4 performed rather well on embedded platforms[51)]. Due to its higher speed and accuracy, the YOLOv4 model is our top pick for large data sets, beating out the Faster R-CNN, YOLOv3, and YOLOv5 models. By switching to CSPDarknet53 as the foundation of the feature extractors in YOLOv4, the algorithm's speed and accuracy have been greatly enhanced.

The formula for the confidence score is given in below equation 3:

$$Confidence\ score = Pr(Object) * IoU\ (3)$$

The significance factor for each item is computed after feature extraction and tag addition, by assigning a value of 1 to the Pr(item) for the ground truth box and a value of 0 to the adjoining grid pixels outside of which there are no objects.

5.7. Images and objects' characteristics embedded

The YOLOv4 output part is concatenated into the encoder output as the final series, allowing us to take advantage of both YOLO Darknet's expertise in object recognition and Xception's in classifying images.

The format of this result is (101 x 2048). One single layer of 256 bits is used with full connectivity for embedding. In order to make the retrieved features uniform in size, they are mapped to a minor dimension that the language decoder can work with[52)].

5.8. Attention mode

Humans possess a highly complex mental ability called "attention mechanism." When presented information,

people can choose to ignore certain pieces of primary data in favor of other pieces of secondary data. Recent advances in picture captioning have greatly benefited from neural networks' attention mechanism, which focuses the network on a subset of inputs to select certain features. In this study, researchers employ a technique wherein a whole image is broken down into layers and choices are made depending on certain features inside the image. For instance, given an image and a set of feature encodings, the model will attempt to make a prediction as shown in Figure 19 and Figure 20. Taking into account the image's value, it speculates on what might come next. For example, if the first letter predicted from the image below is "a," the algorithm will look to see what would be the next most essential part of the image. In this case, the word "dog" serves as a focal point for the entire picture. Typical action or other typical parts inside a picture are sought after in the next step, which aims to traverse numerous layers of comprehension after word prediction has been completed[53].

After that, it travels through space without diverting attention from the window's text. The following (lucidchart-designed) cross-modality graphic of Figure 21 describes the extraction and fine-tuning of features using an embedded YOLO object feature model, as well as image and caption pairs supplied to the attention model for visual-based extractions[54,55]. Captions derived from the decoder using the retrieved visual elements of the multimodal architecture are envisioned in the final predictable inference of test data.

The model's attention-based networks would aid in understanding pivotal parts of the image. The model's textual output would be generated by focusing on select regions of an image using the visual attention technique[56]. The hidden state would then be decoded by a GRU, which would be used to create the labels. Each sequence element takes in both newly generated information and the outputs of previously processed elements[57]. Thus, the RNN connections learn to remember information, which could make captions more accurate and insightful.

The decoder in our model creates new words. The decoder iteratively constructs the output sentence's words based on the visual representation produced by the encoder as shown in Figure 22. For any given image I, there exists a mathematical description of the probability distribution p(S|I) over the caption sentence S = w1, w2, w3,... w S. A dataset is used to train the model to maximize the posterior using equation 4.

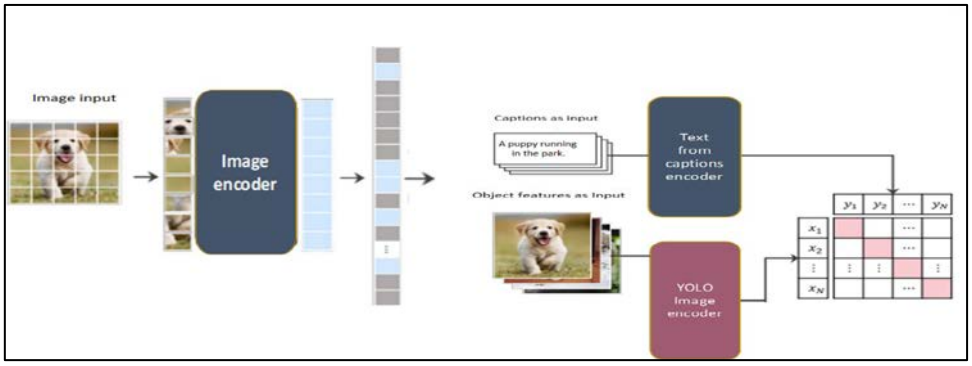

Fig. 19: Multi-modal representation and fine-tuning using text and image characteristics from a pre-trained encoder

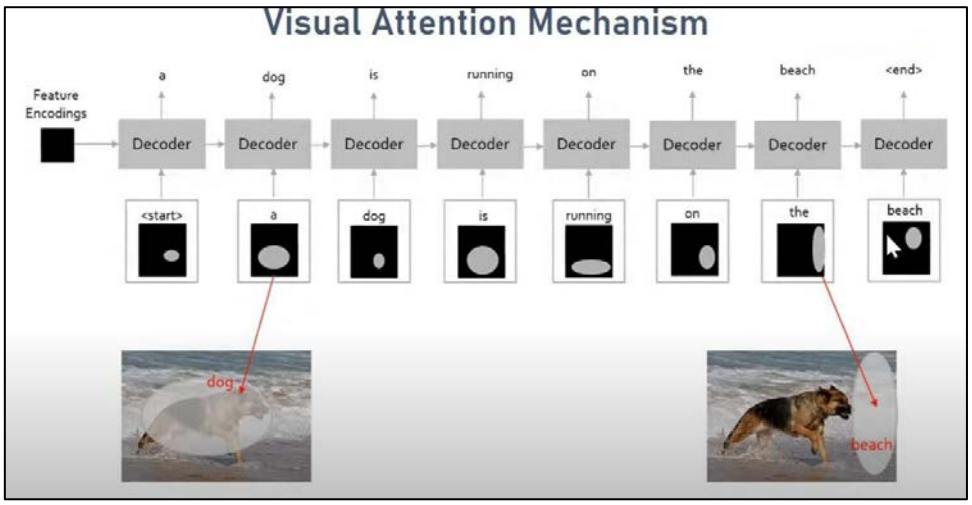

Fig. 20: Visual Attention mechanism of Image

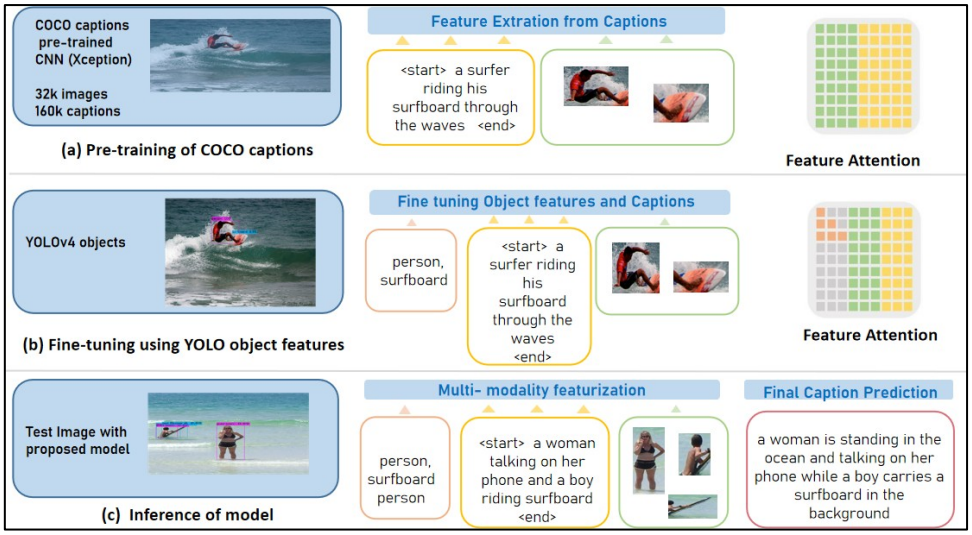

Fig. 21: The recommended method for teaching people how to create captions for images

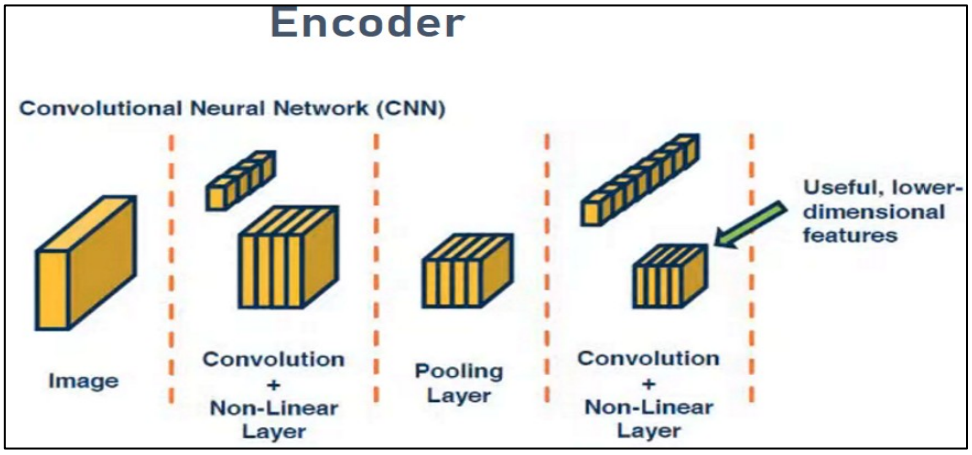

Fig. 22: Diagram of the Vanilla CNN– Encoder

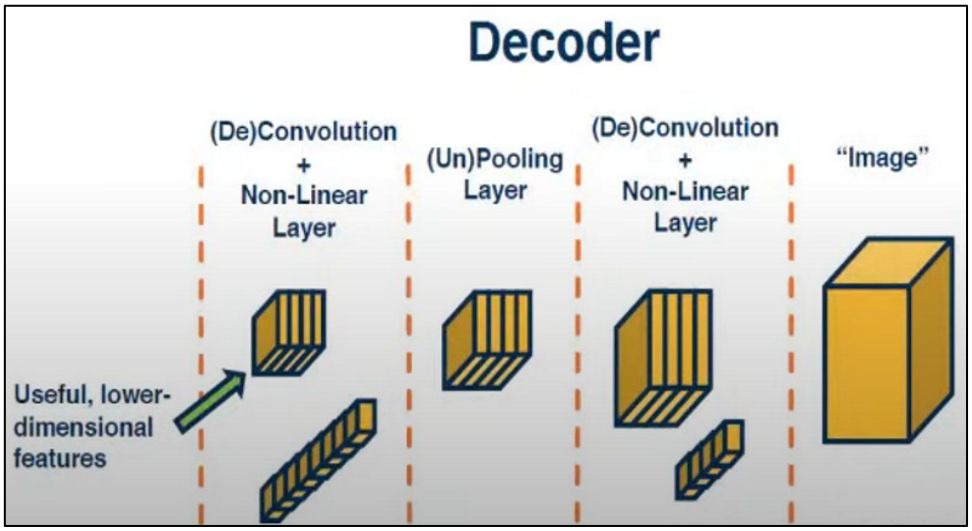

Fig. 23: Vanilla Decoder Architecture Decoder

$$P(w_1, w_2, \ldots, w_S \mid I) = \prod i = 1SP(w_i \mid I, w_1, w_2, \ldots, w_i - 1) \ (4)$$

Each word is supposed to be created in response to the previous word (w1: i 1), considering the distribution.

The GRU used in this article is the RNN unit. This method addresses computing costs and overfitting while offering the most common gated recurrent unit (GRU) alternatives. LSTM is more sophisticated than GRU. Thus, instruction is usually simpler. Despite having more parameters, LSTMs and GRUs have almost completely supplanted vanilla RNNs due to their efficiency and quicker training times. When analyzing the same dataset, GRU trains models 29.29% faster than LSTM. GRU outperforms LSTM only for long text and specialized datasets. GRU wins LSTM with limited training data and long training text, but fails otherwise.

$$\varphi^* = argmax \sum_{(I,t) \in D'} log\ p_\varphi(q|I)$$

$$where\ log\ p_\varphi(q|I) = \sum_{i=1}^{512} \delta[q^i\ =\ Q(t)]\ p_\varphi(q^i|I) \qquad (5)$$

5.9. Classification of Class labeling utilizing POS Tagging Clusterization derived from captions

Clustering is an unsupervised machine learning technique that looks for groupings of items that are quite similar to each other. This strategy enhances the precision of following classification processes in the creation of classifier models[58]. Because our model uses the generated sequence for the sentence probability, we compute the sequence of POS tags before generating an inference. There are several options for obtaining the list of POS tags in sequence[59].

This preliminary method is not scalable. The second method, which we will describe in full below, yielded the most accurate results in our tests.

The classifier is provided with quantified POS tag sequences by cluster medoids. The variable q, denoted as Q(t), represents a sequence of input tags that are represented by the k nearest points in a quantized space[60,61]. The quantization function Q(t) in our terminology transforms t into its quantized tag sequence q. The part-of-speech classifier we employ is trained using the normalized POS series space. This is achieved by incrementing the probability p(q I) to generate projections. As shown in equation 5 , the process of training involves the generation of parameters for both the captioning network and the POS Clustering network. This particular methodology is commonly referred to as Part-of-Speech (POS) tagging. Both of these networks can be trained separately in a direct and uncomplicated manner.

Because the sampled POS sequence q may not correspond to the exact representations of y, noise is produced when there are discrepancies between the two. Each iteration, we draw from a pool of possible POS tag sequences in order to extract the noun tags, and we choose q if and only if it naturally fits in with the POS tags of the caption y.

Class objects occupy 36.3% of these nouns, as shown in Table 4. The decoder's COCO descriptions for the photos say this shows item value. To avoid conflicts, we preconfigured a set of mutually incompatible thing types like COCO item classes. Figure 23 displays our class label hierarchy, with the vehicle class parenting motorbike, airplane, bus, and train and the animal class parenting cats, dogs, horses, and sheep. Annotations as supercategory tags make generic classes the top nodes. Finally, we create classifier models using the cosine similarity of normalized data point vectors and remove context-based lists, clustering them by POS tag cardinality as shown in Figure 24. Displays clusters as Clustset. Since the labels attached to words in the annotated class categories are reliable, we can increase cluster quality. We remove words from clusters without tags.

Table 4: POS Tags at the Item Level

Level	Objects
Captioned foreground elements at the pixel level	63.70%
Noun tags from captions	36.30%

5.10. Training and Test Presumption

The input and output sequences will be generated using 32 thousand training photos. After fitting orders to the model, model.fit_generator() saves the model to a folder. The decoder begins by reading the output of the final concealed state's CNN and YOLOv4 encoders. With x1 as the START> vector and y1 as the first word, we encode the object's features and set up encoder-generated labels.In this case, the network is expected to anticipate the second word, therefore the first word ends with x2=, and the final word, xt, is followed by the END> token, yt. During training, the decoder always receives the proper input, regardless of any previous mistakes.

The above image shows the planned building. The model creates the hidden state sequence $(\mathrm{h}_1, \mathrm{h}_2, \mathrm{h}_3, \ldots \mathrm{h}_\mathrm{n})$ using the image pixel sequence I and the input word vector sequence $(x_1, x_2, x_3, \ldots x_n)$ to produce the output sequence $(y_1, y_2, y_3, \ldots y_n)$. Image feature vectors are only communicated once in their initial, concealed state as shown in Figure 25. To this end, the next hidden state is calculated using the current input x_t, the previously computed image vector I, and the previously computed previous hidden state h_{t-1}. Furthermore, given the input photos, the resulting captions or labels are encoded with the highest feasible log-likelihood. The overall shape of the function is given by the following equation 6:

Cite: S. Bhandari et al., "Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies". Evergreen, 12 (04) 1840-1866 (2025). https://doi.org/10.5109/7402620.

$$L = \sum_{I,S\in X} \sum_{i=1}^{|S|} log\ P(w_i|w_{1:i-1}|I,\theta)\ (6)$$

In this case, S is a caption sentence $\{w_1, w_2, w_3, \ldots w_{|S|}\}$ from the X training dataset, and w1, w2, w3,...w is the parameter value. The conditional probability of the current word $w_t$is X; $P(w_i|w_{1:i-1}|I;\theta)$ if we consider the input picture I, the model parameters, and the past words created $w_{1:i-1}$throughout the evaluation stage, we do not have access to the whole caption like we did throughout training. Therefore, an iterative approach is used. The word generator selects a random sample of words from the target distribution and uses those words to seed the subsequent time step. When the special token End> is accessed for the last time, the iterative process stops.

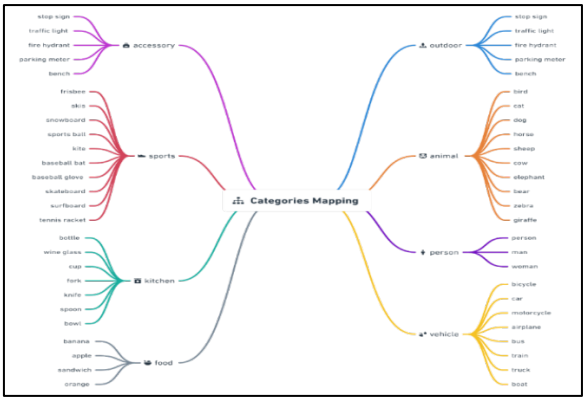

Fig. 24: Class and Super Category Hierarchy for Easier Cardinality Measurement

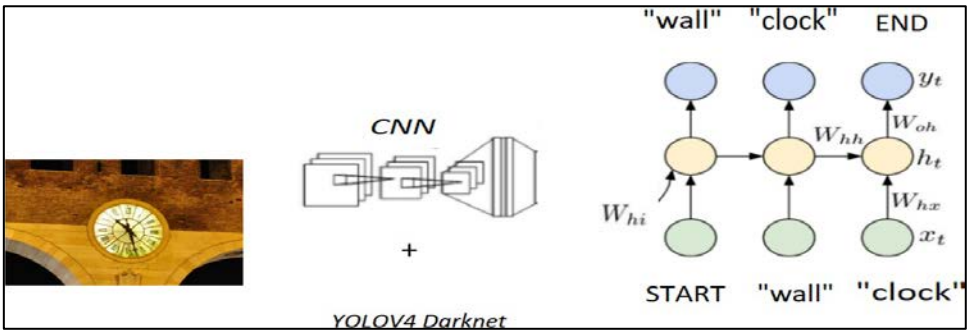

Fig. 25: Image-text description generation architecture

A straightforward approach for context-aware POS labeling of a list is in[62,63]. We identify words in an example sentence based on cluster frequency and proximity to the beginning and conclusion. The cluster tag replaces the test word tag if only one cluster contains the term. When the test word appears frequently in many clusters, the supplemental context word pair is prioritized. Research demonstrates that context word pairs in clusters must start with the test word's comparable prefix. For test phrases that don't fit into any of the classes, we apply these criteria to find helpful clusters. One criterion considers preferred order to choose relevant clusters. ProbLabel, the ratio of greatest to second-highest cluster tag probabilities, helps.

$$Diff\ Prob = LabelProb(C_{max1}) - LabelProb(C_{max2}) / LabelProb(C_{max1})\ (7)$$

where C_{max1} *and* C_{max2} are the highest and second highest *LabelProb* (C_i)value re- 837
spectively where *LabelProb* (C_i) of a cluster is represented as follows:

$$LabelProb(C) = Frequency\ of\ X\ in\ (C) / \sum \forall Cj \in ClustsetFrequency\ of\ X\ in\ (Cj)\ (8)$$

If the test word is in Ci, then the test is at X. If the clusters are chosen based on whether or not they contain the words that serve as context for the test word, then X equals the words that serve as context;
If clusters are selected using the word pair immediately around the test word, then X is those two words; otherwise, X is the word immediately before the test word. We can tag unknown or odd phrases left out of the training set using this alternative smoothing method. Clusters are selected by relevance, and DiffProb(using equation 7), LabelProb(using equation 8) values are produced. If DiffProb is less than probDiffMin, "NOLABEL" is returned and the high LabelProb cluster tag labels the test word. Consider the cat-dog image below, as the model will generate words for each test image I. "a dog sitting near a window" and "a cat sitting on a table" are right, although the latter is lacking a crucial component.
To choose the most suitable sentence matching the image, ranking the generated sentences is necessary. The TF-IDF method is employed, and the best candidate is determined based on the highest scores among the generated sentences. The ranking involves clustering noun phrases using weight computations, where the frequency of POS n-grams, denoted as (p_f), is calculated (as shown in equation 9) by counting the occurrences of the term in the POS n-grams:

$$log\ \frac{|C|}{(p_f)} \quad (9)$$

The value |C| represents the total number of POS noun phrase collections. This value aids in ranking the system's selection of the phrase most similar to the sample image.

6. Results and Discussion

We share generational and comparative findings from our methodology's several phases of evaluation. Using the MS COCO 2014 dataset, we have trained the relevant algorithms, gathering data from both the validation and test sets with a maximum of five iterations per group to prevent overfitting on the test set.

COCO metrics were evaluated with the COCOEvalCap API, while Clustering and Multi-class classification were evaluated with Scikit-Learn.

6.1. Individual and Embedded Evaluation results on Captions

Through the utilization of BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), BLEU-4 (B-4), CIDEr, METEOR, ROUGE-L, and SPICE, we assess our findings on the COCO dataset's Encoder-Decoder model. The next step is to apply Accuracy, Confidence, Precision, and Recall metrics to the outcomes of POS label tagging of Categories utilizing clusterization and then genuine categorization.

6.2. Encoder-Decoder Model Evaluation

Using the Microsoft COCO dataset and the TensorFlow library, we show our findings. The original data set was split into test and validation collections at the 20% level. The model was trained with a loss function of sparse categorical cross entropy over the course of 30 iterations. Adam was put to use all through the process of optimization. The performance of the proposed model on the MS COCO test data is compared to that of the baseline model improved with Xception characteristics in Table 5. When object features are accounted for in the evaluation process, test results improve dramatically; the CIDEr score, in particular, rises by 37.93%.

Improved grammatical integrity, more salience, and a close correlation with human judgment are all results of using a wide range of object attributes. It indicates that the CIDEr measures grow significantly, when the significance factor is present, the BLEU measures appear to expand in tandem with other metrics.

6.3. Comparison of Results with Baseline

Our primary cause for outcomes improvement is compared to MS COCO's in Table 6[67]. Image captions were used to test item properties. To get object properties, we encoded the YOLO model's object configurations with an LSTM.After obtaining the VGG16 CNN features, a separate LSTM unit encodes and appends them to the feature vectors. Replace YOLO with CNN for object characteristics.5,000 example photographs test a dataset subset[68]. For each criterion, boldface represents the largest score difference. Encoding and integrating CNN and object feature types independently gives them equal weight in the baseline model. Our results may differ since their benchmark is more dependable. B-1, B-2, and B-3 used different SPICE and BLEU scores. The model's SPICE grew by 23.14 percent after adding object tags, demonstrating semantic link. SPICE is difficult to change. Feature combination and encoding approach may affect baseline and model scores validation and testing sets are qualitatively compared in Table 6[69].

The model operates by first identifying the various objects present in Figures 26, 27, and 28. Subsequently, it engages in the process of generating descriptive captions corresponding to the identified objects. Notably, the bounding boxes outlining the spatial extents of these objects, along with their respective labels, are concurrently displayed in conjunction with the generated captions. This

Table 5: Embedding of object features vs the isolated baseline model, along with a comparison of their significant factors and relative improvements in assessment scores

Model	B-1	B-2	B-3	B-4	CIDEr	METEOR	SPICE	ROUGE-L
Captioning a single CNN image	0.492	0.296	0.174	0.101	0.435	0.163	0.108	0.358
Boundary boxes based on the YOLO theory	0.501	0.355	0.237	0.162	0.571	0.183	0.133	0.367
Model using YOLO limits and the impact measure	0.579	0.404	0.279	0.191	0.6	0.195	0.133	0.396
Rise in percentage after embedding of object features	17.68	36.48	60.34	89.1	37.93	19.63	23.14	10.61

integrated approach ensures a comprehensive understanding and representation of the visual content within the image[70,71]. Images comparing the baseline model with the enriched model (which takes object attributes into account) are shown in Figures 28, 29, and 30. All evaluation metrics corroborate our hypothesis that accuracy can be improved by incorporating features for object recognition into the vision model[70,71].

Also, in Figure 31 below showcase the instances of randomized images along with their consensus captions. Captions selected with CIDEr exhibit slightly have greater comprehensiveness. These consensus captions are generated using the scoring metrics CIDEr and BLEU.

6.4. Results of POS Labeling Strategy

In the Parts-of-Speech Evaluation, outlined in Table 7, our analysis involved 5,000 test images and a set of 25,549 items within the embedded model of the test data. These observations were categorized into k clusters through K-means clustering. It's important to note that the classification process is independent of the number of classes, as the number of clusters formed is distinct from class categorization.

Table 6: Comparison with the metrics

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
(Yin & Ordonez, 2017) results with object features	NA	NA	NA	0.253	0.238	0.922	0.507	NA
Our model with YOLO bounding boxes and the significance factor	0.579	0.404	0.279	0.191	0.195	0.6	0.396	0.133

Fig. 26: Model recognizes objects: elephants, persons, and generates captions for an image, with bounding boxes and labels captured concurrently

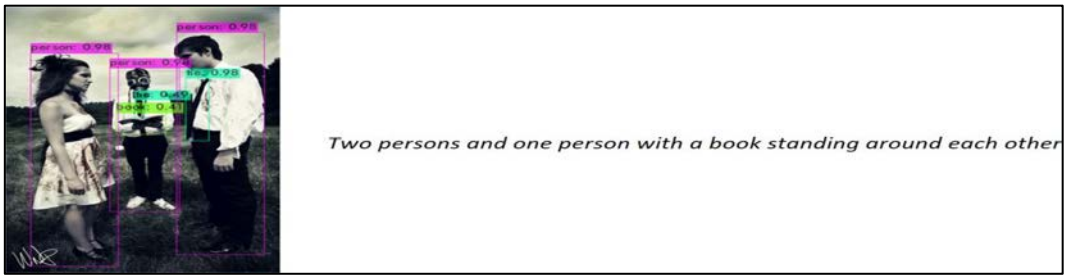

Fig. 27: Similarly, model objects: book, person, tie and generates captions for an image, with bounding boxes and labels captured concurrently

Fig. 28: Without object features, a man in the image is erroneously labelled as the model plan

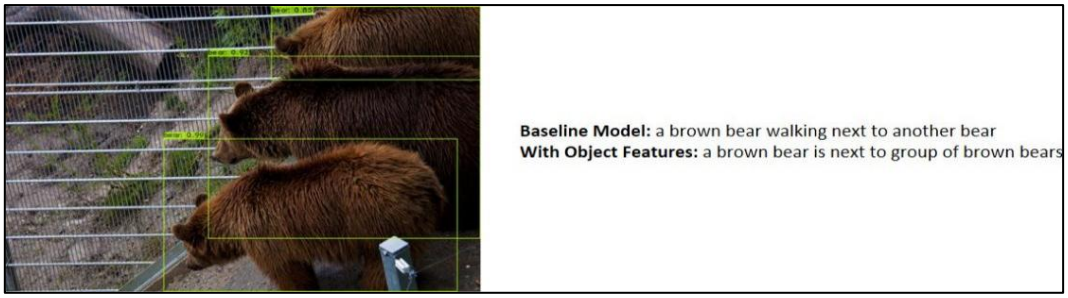

Fig. 29: The model fails to recognize the third bear due to a lack of essential object features inclusion, but when equipped with these features, it accurately generates captions

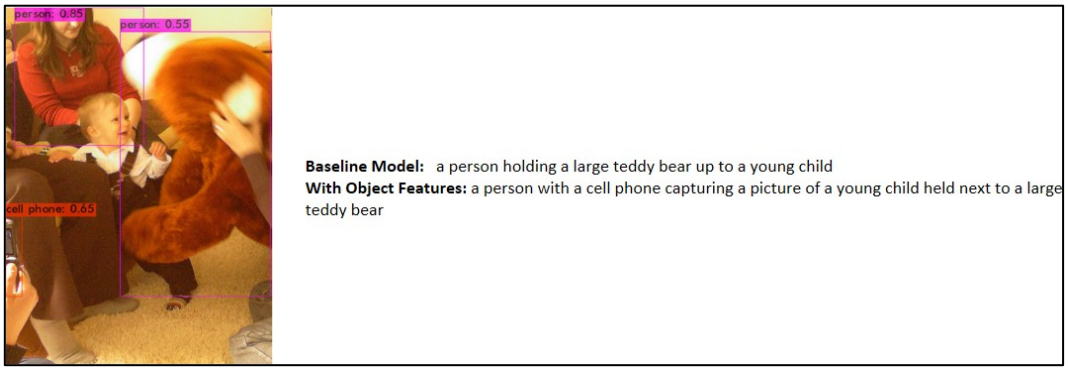

Fig. 30: The model includes the object feature of a cell phone in the detected captions

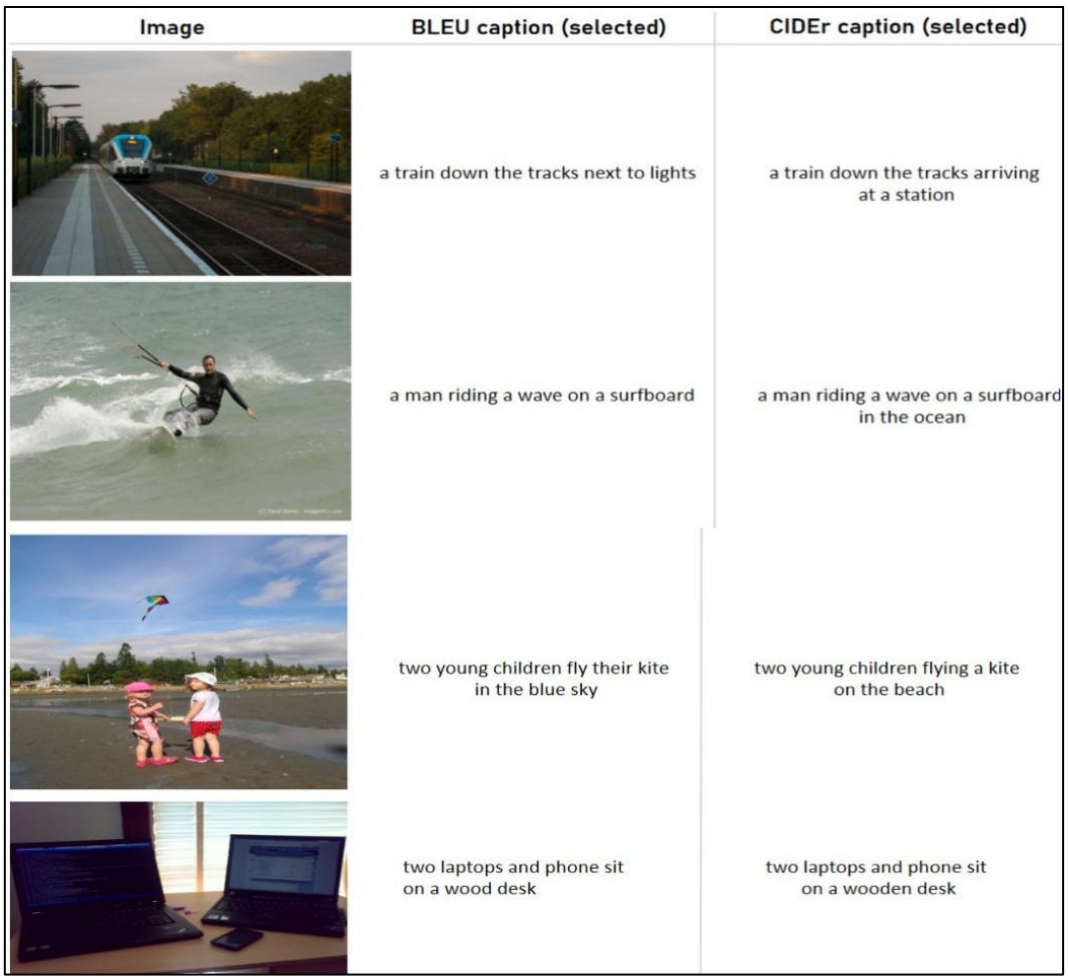

Fig. 31: Comparison of captions between BLEU and CIEDr

Table 7: See how the 22,352 words in the test set are affected by the size of the captions data below using the classifier model

No. of words in Captions set	No. of clusters in model	No. of NOLABEL test words	Average accuracy (%)
25549	12	2647	82.54

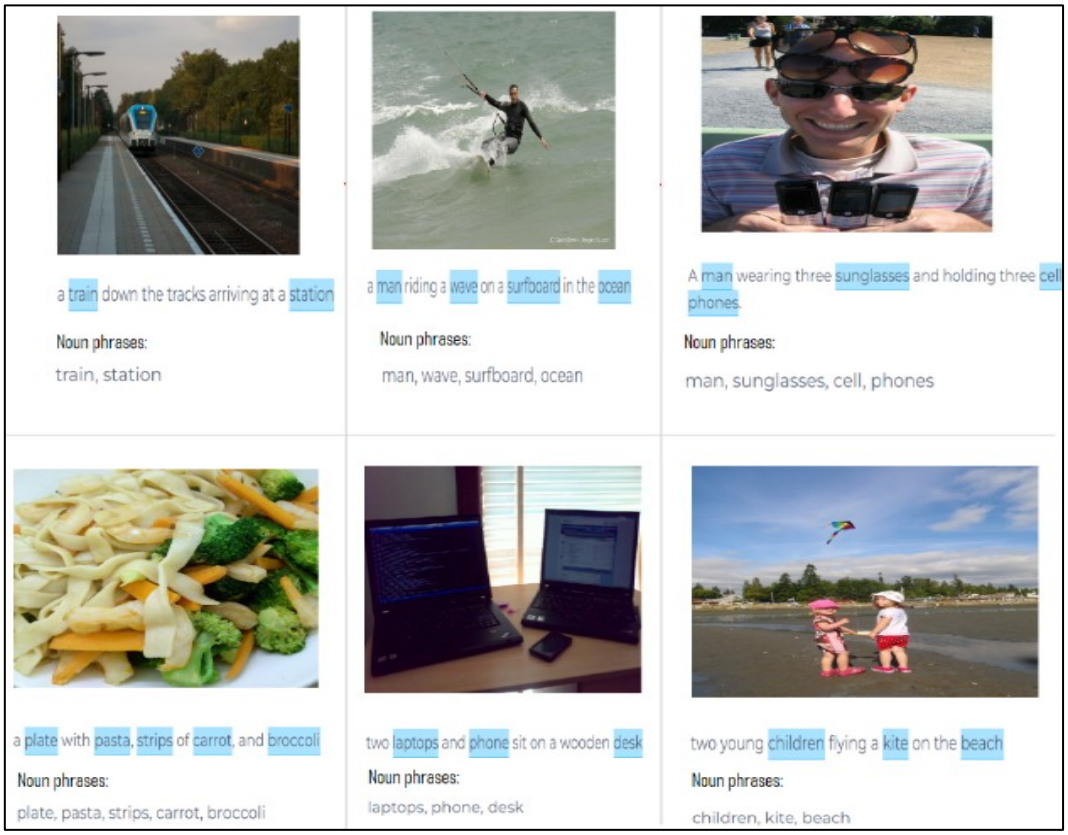

Fig. 32: Shows how the suggested model's use of object features can be used to extract Nouns from captions. The blue-highlighted words in the embedded object feature captions created by the Encoder-Decoder system denote nouns

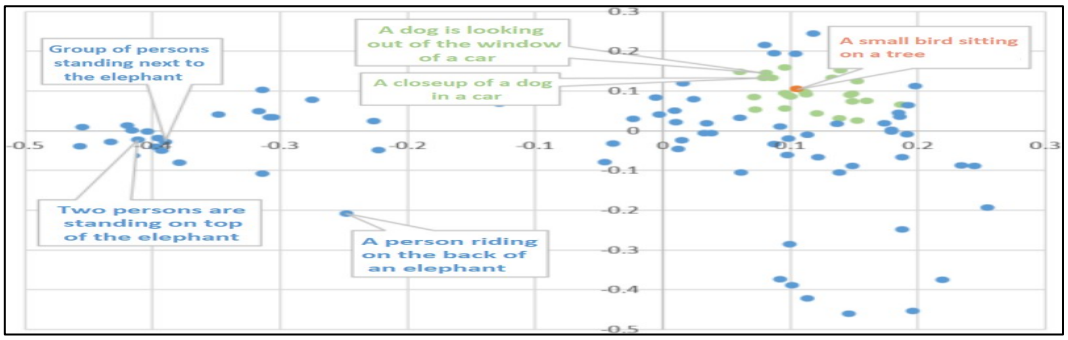

Fig. 33: The image above is an illustration of a set that could be clustered. An idealized categorization procedure, represented in two dimensions by the highest-scoring (green) and consensus (orange) captions

First, we train a simple KNN classification model with 75.43 percent success. Our model using K Means clusters had 82.5% accuracy after label classification. With these test data definitions, the optimal values for Confidence, Coverage, and DiffProb are AverageAccuracy, MinConfidence = 70%, and MinprobDif = 30%.

In Figure 32, the depicted results showcase the extraction of nouns from captions generated by the proposed model, utilizing object features. The words highlighted in blue specifically represent the nouns extracted from the captions generated through the Encoder-Decoder, which is enriched with object features.

Figure 33 shows hypothetical clustering with a restricted amount of captions to depict a theoretical classification

process in which the captions set for the 30,500 test set words was altered. Orange is the consensus caption, whereas green represents the highest-scoring caption. With more annotated testing data, the classifier model's word cluster penetration (number of unique terms in the cluster set) improved. Because lists can be contextually categorized better. Include identifiers in a cluster to increase the number of distinct terms.

We employed a dimensional reduction technique to condense the embedding space into three dimensions. Utilizing weighted averaging for supercategory representation, we emphasized images belonging to the same taxonomy through PCA grouping of TF-IDF. Following the reduction of the embedding space to three dimensions, we applied PCA clustering of TF-IDF to accentuate the similarities among images.

The categorical classification signifies model's accuracy and recall for each predicted class. Precision and recall are best for people, cats, horses, and dogs, and poorest for a microwave oven, a hair dryer, a book, a toothbrush, and a stop sign. Classifications that use the term "person" tend to be the most reliable and accurate as shown in Figure 34 and Figure 35.

7. Real-time Impact of the Proposed Work

In the pursuit of productivity, leveraging time effectively is a universal aspiration. It's not about having more time but about transforming it into meaningful accomplishments. Success often lies in the meticulous documentation of thoughts, ideas, and tasks, a practice favoured by high achievers.

Consider a scenario where an individual accumulates a repository of Ayurvedic remedies for various ailments, organized through machine learning techniques. These remedies, sourced from the ancient wisdom of Ayurveda, are tagged with labels based on symptoms and ingredients. The individual, when faced with an ailment, can swiftly access the relevant remedy, thanks to meticulous data organization.

The Machine Learning Intervention: From Data Chaos to Practical Wisdom:

In a world inundated with data, effective organization becomes paramount. Machine learning, particularly through the lens of generative processes and cross-modal feature embedding, presents a transformative solution. This research delves into the integration of these techniques, highlighting their prowess in segregating and categorizing diverse data forms.

The core aim is to bridge the gap between human annotations and extracted features, utilizing word objects to categorize labels. The research aspires to promptly generate natural language captions, providing a producible target language that accurately describes image elements.

At the heart of this research is the belief that data, when organized effectively, becomes a powerful tool for knowledge and decision-making. The motivation stems from the recognition that in the era of abundant data, the key lies in extracting meaningful information and making informed choices.

Empowering Time and Knowledge Management:

Consequently, this research not only addresses the challenges of organizing massive datasets but also offers a glimpse into a future where machine learning aids in swift and intelligent decision-making. The case study presented exemplifies the practical impact of these methodologies, demonstrating the seamless integration of ancient wisdom with modern technological advancements.

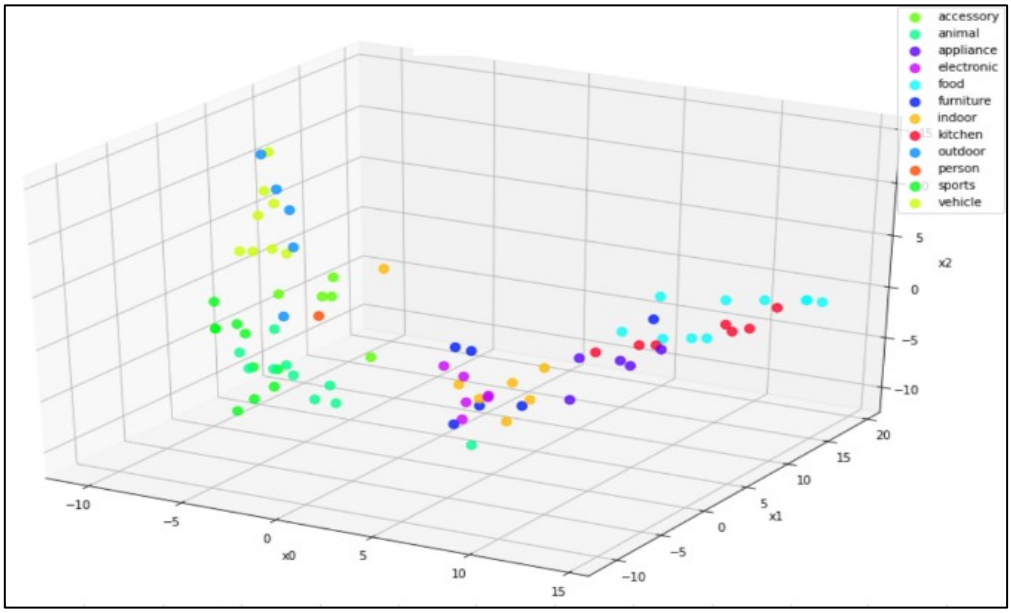

Fig. 34: K-means clustering of classes based on supercategory with assigned weightage

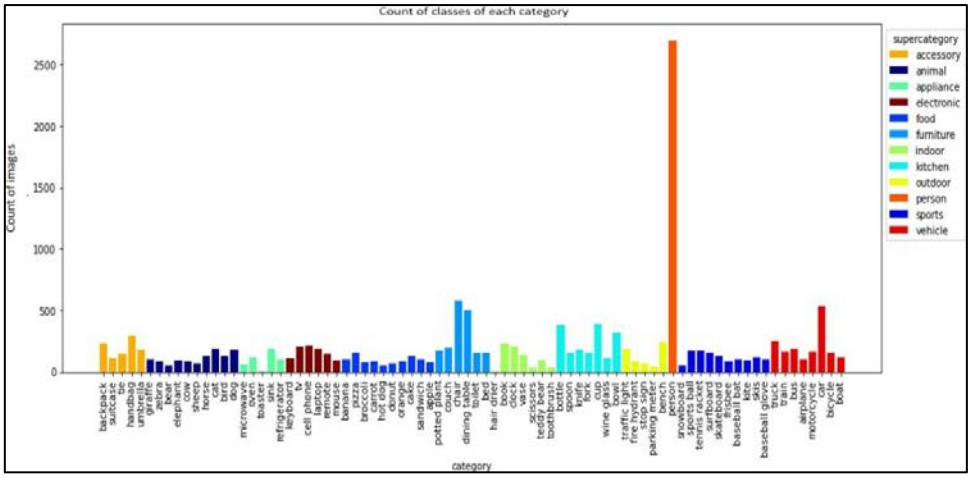

Fig. 35: Count of test dataset images categorized by detected super categories for class categories

8. Conclusion and Future Work

We employed CNN (Xception)-pretrained image classification features and YOLOv4 feature module object detection to establish that an attention-based Encoder-Decoder model for image captioning works. The significance factor gives larger, more confident items and those in the foreground more weight than smaller, less confident ones in the background. Deep neural networks have helped this sector make headway despite the problem's challenges. Most research use more advanced CNN and RNN for encoding-decoding. In these tests, the input image is encoded using a fixed-length CNN feature vector, which is then fed into a recurrent neural network (RNN) that conditions the output word at each time step based on the input visual vector and the words created in the previous time steps.

The encoded visual vector from the whole image lacks a temporal idea, which is essential to RNN memory, which may diminish its visual information memory. The statistical language models we utilize predict how phrases will be used to generate new sentences.

We demonstrate that phrase production can be solved without sophisticated recurrent networks. CNN-level features and YOLO object properties generate new words, which our final model clusters into a few super categories. Our model may gain from merging CNN visual information with text characteristics from photo counts and location metadata. Possible enlargements are being saved for investigation.

The potential limitations of this work are the use of the MS COCO dataset, and this is likely to reduce its scope of applications in other areas. It makes use of pre-set object classes, where flexibility towards unseen objects is restricted. The model has the risk of performance degrading on poorly quality images and the computational costs of the model can grow as the models become more complex or as the datasets become larger in size.

Future work will improve phrase creation with more complicated language models and supervised data. Visually grounded phrase representations may alter natural language processing algorithms, which is exciting.

Improving convolutional feature-based object detection can advance this field. Using large-scale NLP corpora, the Pytesseract OCR paradigm may create Custom objects that embed and map to COCO objects in addition to object features. Future image categorization can benefit from caption texts and object layouts' rich semantic data. Modern methods encrypt and send object characteristics to decoders. Complex language models like Meshed-Memory Transformers can now contribute captioning terms to this taxonomy, previously only available for COCO. Use Open Images, Flickr, or another dataset to teach your model to recognize many items and distinguish between foreground and background. The high confidence score lets us generate several words in natural language for consumers to better grasp mark localization. The multi-fold model should be expanded to incorporate the taxonomy label suggestion for image and text attributes from embedded photos.

In future, it is possible to research into incorporation of more advanced transformer-based language models such as BERT or GPT so as to provide more fluency and context enriched captioning. Domain adaptation and transfer learning can be augmented to be more robust across a wide range of domains that could be medical or satellite imagery. Scalability to caption generation on multiple languages increases the accessibility of the world. Object recognition can be improved by introducing OCR and context-relevant embeddings. Edge systems provide useful possibilities such as smart surveillance in real time. Larger data sets and multi-benchmark assessments are able to ameliorate generalization and expose biases. Improving explainability of attention mechanisms leads to increases in level of trust and usability. Finally, audio or sensor sensorimotor data integration as multi-modal reasoning may enrich both the interpretation of the scene and storytelling.

Cite: S. Bhandari et al., "Hybrid Vision-and-Language Fusion: A Threefold Learning Approach for elevating Image Captioning through Adaptive Strategies". Evergreen, 12 (04) 1840-1866 (2025). https://doi.org/10.5109/7402620.

References

1) S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 1508–1517 (2020). doi:10.1109/WACV45572.2020.9093430.
2) Y. Wang, Y. Xie, J. Zeng, H. Wang, L. Fan, and Y. Song, "Cross-modal fusion for multi-label image classification with attention mechanism," *Comput. Electr. Eng.*, 101, 108002 (2022). doi:10.1016/j.compeleceng.2022.108002.
3) Y. Xie, Y. Wang, Y. Liu, and K. Zhou, "Label graph learning for multi-label image recognition with cross-modal fusion," *Multimed. Tools Appl.*, 1–19 (2022). doi:10.1007/s11042-022-12397-y.
4) X. Xue, and J. Zhang, "Part-of-speech tagging of building codes empowered by deep learning and transformational rules," *Adv. Eng. Inform.*, 47, 101235 (2021). doi:10.1016/j.aei.2020.101235.
5) J. Yang, Y. Sun, J. Liang, B. Ren, and S.-H. Lai, "Image captioning by incorporating affective concepts learned from both visual and textual components," *Neurocomputing*, 328, 56–68 (2019). doi:10.1016/j.neucom.2018.03.078.
6) X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10685–10694 (2019). doi:10.1109/CVPR.2019.01094.
7) K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2det: Learning to amplify weak caption supervision for object detection," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 9686–9695 (2019). doi:10.1109/ICCV.2019.00978.
8) X. Yin, and V. Ordonez, "Obj2text: Generating visually descriptive language from object layouts," *arXiv preprint arXiv:1707.07102* (2017). doi:10.18653/v1/D17-1017.
9) J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, 30 (12), 4467–4480 (2019). doi:10.1109/TCSVT.2019.2917468.
10) F. Zhu, Z. Ma, X. Li, G. Chen, J.-T. Chien, J.-H. Xue, and J. Guo, "Image-text dual neural network with decision strategy for small-sample image classification," *Neurocomputing*, 328, 182–188 (2019). doi:10.1016/j.neucom.2018.10.100.
11) J. Aiswarya, K. Veerappan, and K. Mariammal, "Categorization of on-road automobiles using deep learning approach," *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, 227–233 (2022). doi:10.1109/ICSCDS53736.2022.9761036.
12) M.A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *J. Big Data*, 9 (1), 1–16 (2022). doi:10.1186/s40537-022-00571-w.
13) M.D.S. Alam, M.D.S. Rahman, M.D.I. Hosen, K.A. Mubin, S. Hossen, and M.F. Mridha, "Bahdanau attention-based Bengali image caption generation," *Proc. Int. Conf. Decision Aid Sci. Appl. (DASA)*, 1073–1077 (2022). doi:10.1109/DASA54658.2022.9765268.
14) H.N. Alkalouti, and M.A.A.L. Masre, "Encoder-decoder model for automatic video captioning using YOLO algorithm," *Proc. IEEE Int. IoT Electron. Mechatronics Conf. (IEMTRONICS)*, 1–4 (2021). doi:10.1109/IEMTRONICS52119.2021.9422600.
15) P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 6077–6086 (2018). doi:10.1109/CVPR.2018.00636.
16) J. Aneja, H. Agrawal, D. Batra, and A. Schwing, "Sequential latent spaces for modeling the intention during diverse image captioning," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 4261–4270 (2019). doi:10.1109/ICCV.2019.00436.
17) D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473* (2014).
18) Y. Bao, M. Wu, S. Chang, and R. Barzilay, "Few-shot text classification with distributional signatures," *arXiv preprint arXiv:1908.06039* (2019).
19) A. Bochkovskiy, C.-Y. Wang, and H.-Y.M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934* (2020).
20) J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient adaptation of pretrained language models for image captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 18030–18040 (2022). doi:10.1109/CVPR52688.2022.01753.
21) T.H. Chen, Y.H. Liao, C.Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 521–530 (2017). doi:10.1109/ICCV.2017.62.
22) Y.C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Universal image-text representation learning," *Lect. Notes Comput. Sci.*, 12375, 104–120 (2020). doi:10.1007/978-3-030-58577-8_7.
23) K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, 17

(11), 1875–1886 (2015). doi:10.1109/TMM.2015.2477044.

24) F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1251–1258 (2017). doi:10.1109/CVPR.2017.195.
25) S. Chun, W. Kim, S. Park, M. Chang, and S.J. Oh, "ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO," *arXiv preprint arXiv:2204.03359* (2022).
26) M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 10578–10587 (2020). doi:10.1109/CVPR42600.2020.01059.
27) J. Fan, "OPE-HCA: An optimal probabilistic estimation approach for hierarchical clustering algorithm," *Neural Comput. Appl.*, 31 (7), 2095–2105 (2019). doi:10.1007/s00521-015-1998-5.
28) A.K. Gangwar, and V. Ravi, "A novel BGCapsule network for text classification," *SN Comput. Sci.*, 3 (1), 1–12 (2022). doi:10.1007/s42979-021-00916-0.
29) P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 5825–5835 (2019). doi:10.1109/ICCV.2019.00592.
30) J. Gu, J. Cai, S.R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 7181–7189 (2018). doi:10.1109/CVPR.2018.00750.
31) L. Gui, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Training vision-language transformers from captions alone," *arXiv preprint arXiv:2205.09256* (2022). doi:10.48550/arXiv.2205.09256.
32) Y. Hirota, N. Garcia, M. Otani, C. Chu, Y. Nakashima, I. Taniguchi, and T. Onoye, "A picture may be worth a hundred words for visual question answering," *arXiv preprint arXiv:2106.13445* (2021). doi:10.48550/arXiv.2106.13445.
33) I. Hrga, and M. Ivašić-Kos, "Deep image captioning: An overview," *Proc. 42nd Int. Conv. Inf. Commun. Technol. Electron. Microelectron. (MIPRO)*, 995–1000 (2019). doi:10.23919/MIPRO.2019.8756680.
34) L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 4634–4643 (2019). doi:10.1109/ICCV.2019.00474.
35) F. Jánez-Martino, E. Fidalgo, S. González-Martínez, and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning," *arXiv preprint arXiv:2005.08773* (2020).
36) P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Procedia Comput. Sci.*, 199, 1066–1073 (2022). doi:10.1016/j.procs.2022.01.135.
37) Y. Jin, Y. Chen, L. Wang, J. Wang, P. Yu, L. Liang, J.-N. Hwang, and Z. Liu, "The overlooked classifier in human-object interaction recognition," *arXiv preprint arXiv:2203.05676* (2022).
38) G.C. Kang, S. Kim, J.-H. Kim, D. Kwak, and B.-T. Zhang, "The dialog must go on: Improving visual dialog via generative self-training," *arXiv preprint arXiv:2205.12502* (2022).
39) R. Khan, M.S. Islam, K. Kanwal, M. Iqbal, M. Hossain, and Z. Ye, "A deep neural framework for image caption generation using GRU-based attention mechanism," *arXiv preprint arXiv:2203.01594* (2022).
40) D.J. Kim, J. Choi, T.-H. Oh, and I.S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 6271–6280 (2019). doi:10.1109/CVPR.2019.00642.
41) J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 16478–16488 (2021). doi:10.1109/CVPR46437.2021.01620.
42) P. Li, P. Chen, Y. Xie, and D. Zhang, "Bi-modal learning with channel-wise attention for multi-label image classification," *IEEE Access*, 8, 9965–9977 (2020). doi:10.1109/ACCESS.2020.2965110.
43) J. Lin, A. Yang, Y. Zhang, J. Liu, J. Zhou, and H. Yang, "InterBERT: Vision-and-language interaction for multi-modal pretraining," *arXiv preprint arXiv:2003.13198* (2020).
44) T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO: Common objects in context," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 740–755 (2014). doi:10.1007/978-3-319-10602-1_48.
45) C. Liu, C. Wang, F. Sun, and Y. Rui, "Image2Text: A multimodal caption generator," *Proc. ACM Multimedia*, 746–748 (2016). doi:10.1145/2964284.2973831.
46) P. López-Úbeda, M.C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L.A. Ureña-López, and M.T. Martín-Valdivia, "Automatic medical protocol classification using machine learning approaches," *Comput. Methods Programs Biomed.*, 200, 105939 (2021). doi:10.1016/j.cmpb.2021.105939.
47) Q. Mao, C. Wang, S. Yu, Y. Zheng, and Y. Li, "Zero-shot object detection with attributes-based category similarity," *IEEE Trans. Circuits Syst. II Express Briefs*, 67 (5), 921–925 (2020). doi:10.1109/TCSII.2019.2959072.
48) E. Merdivan, A. Vafeiadis, D. Kalatzis, S. Hanke, J.

Kroph, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Image-based text classification using 2D convolutional neural networks," *Proc. IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IoP/SCI*, 144–149 (2019). doi:10.1109/SmartWorld-UIC-ATC-ScalCom-CBDCom-IoP.2019.00045.

49) M. Muscetti, A.M. Rinaldi, C. Russo, and C. Tommasino, "Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques," *Knowl. Inf. Syst.*, 64 (5), 1283–1303 (2022). doi:10.1007/s10115-022-01515-0.

50) U. Nepal, and H. Eslamiat, "Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs," *Sensors*, 22 (2), 464 (2022). doi:10.3390/s22020464.

51) C.C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 41 (4), 999–1012 (2018). doi:10.1109/TPAMI.2018.2797607.

52) J. Prudviraj, C. Vishnu, and C.K. Mohan, "M-FFN: multi-scale feature fusion network for image captioning," *Appl. Intell.*, 1–13 (2022). doi:10.1007/s10462-022-10047-0.

53) A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 8748–8763 (2021). doi:10.48550/arXiv.2103.00020.

54) P. Rani, V. Pudi, and D.M. Sharma, "A semi-supervised associative classification method for POS tagging," *Int. J. Data Sci. Anal.*, 1 (2), 123–136 (2016). doi:10.1007/s41060-016-0041-0.

55) S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 28 (2015).

56) S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 7008–7024 (2017). doi:10.1109/CVPR.2017.749.

57) A. Sabir, F. Moreno-Noguer, and L. Padró, "Textual visual semantic dataset for text spotting," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 542–543 (2020). doi:10.1109/CVPRW50498.2020.00088.

58) M. Seshadri, M. Srikanth, and M. Belov, "Image to language understanding: captioning approach," *arXiv preprint arXiv:2002.09536* (2020). doi:10.48550/arXiv.2002.09536.

59) H. Sharma, M. Agrahari, S.K. Singh, M. Firoj, and R.K. Mishra, "Image captioning: a comprehensive survey," *Proc. Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, 325–328 (2020). doi:10.1109/PARC48935.2020.00075.

60) D. Sileo, "Visual grounding strategies for text-only natural language processing," *arXiv preprint arXiv:2103.13942* (2021). doi:10.48550/arXiv.2103.13942.

61) M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: a survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.* (2022). doi:10.1109/TPAMI.2022.3140191.

62) W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530* (2019). doi:10.48550/arXiv.1908.08530.

63) Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *arXiv preprint arXiv:2205.02655* (2022). doi:10.48550/arXiv.2205.02655.

64) D. Wang, J. Wang, F. Hu, L. Li, and X. Zhang, "A locally adaptive multi-label k-nearest neighbor algorithm," *Proc. Pacific-Asia Conf. Knowl. Discov. Data Min. (PAKDD)*, Springer, 81–93 (2018). doi:10.1007/978-3-319-93037-4_7.

65) J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognit.*, 98, 107075 (2020). doi:10.1016/j.patcog.2019.107075.

66) J. Park, and B. Han, "Multi-modal representation learning with text-driven soft masks," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2798–2807 (2023). doi:10.1109/CVPR52729.2023.00274.

67) Y. Ren, Z. Mao, S. Fang, Y. Lu, T. He, H. Du, Y. Zhang, and W. Ouyang, "Crossing the gap: Domain generalization for image captioning," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2871–2880 (2023). doi:10.1109/CVPR52729.2023.00281.

68) I.K. Salman Al-Tameemi, M.R. Feizi-Derakhshi, S. Pashazadeh, and M. Asadpour, "Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data," *IEEE Access*, 11 (7), 91060–91081 (2023). doi:10.1109/ACCESS.2023.3307716.

69) R. Chahar, A.K. Dubey, and S.K. Narang, "A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection," *Int. J. Adv. Technol. Eng. Explor.*, 8 (83), 1279 (2021). doi:10.19101/IJATEE.2021.874198.

70) A. Dubey, U. Gupta, and S. Jain, “Medical data clustering and classification using TLBO and machine learning algorithms,” *Comput. Mater. Contin.*, 70 (3), 4523–4543 (2021). doi:10.32604/cmc.2022.021148.

71) Y. Xu, M. Zhang, X. Yang, and C. Xu, “Exploring multi-modal contextual knowledge for open-vocabulary object detection,” *arXiv preprint arXiv:2308.15846* 14 (8), 1–12 (2023). doi:10.1109/TIP.2024.3485518.