Semantic Context and Attention-driven Framework for Predicting Visual Description Utilizing a Deep Neural Network and Natural Language Processing

K. Annapoorneshwari Shetty [1] & Subrahmanya Bhat [2]

[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,
Orcid ID: 0000-0003-1900-256X; Email: annapoorna.nikesh@gmail.com

[2] Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,
Orcid ID: 0000-0003-2925-1834; Email: sbhat.ccis@srinivasuniversity.edu.in

Area of the Paper: Computer Science.
Type of the Paper: Review Paper.
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed In: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.8187252
Google Scholar Citation: IJCSBE

How to Cite this Paper:
Shetty, K. A., & Bhat, S., (2023). Semantic Context and Attention-driven Framework for Predicting Visual Description Utilizing a Deep Neural Network and Natural Language Processing. *International Journal of Case Studies in Business, IT, and Education (IJCSBE), 7*(3), 119-139. DOI: https://doi.org/10.5281/zenodo.8187252

International Journal of Case Studies in Business, IT and Education (IJCSBE)
A Refereed International Journal of Srinivas University, India.

Crossref DOI: https://doi.org/10.47992/IJCSBE.2581.6942.0290

Paper Submission: 27/05/2023
Paper Publication: 28/07/2023

Semantic Context and Attention-driven Framework for Predicting Visual Description Utilizing a Deep Neural Network and Natural Language Processing

K. Annapoorneshwari Shetty [1] & Subrahmanya Bhat [2]

[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,

Orcid ID: 0000-0003-1900-256X; Email: annapoorna.nikesh@gmail.com

[2] Institute of Computer Science and Information Science, Srinivas University, Mangalore, India,

Orcid ID: 0000-0003-2925-1834; Email: sbhat.ccis@srinivasuniversity.edu.in

ABSTRACT

Background/Purpose: *This literature review's goal is to inspect various machine learning algorithms for visual description and their applications to prediction. Examining the numerous approaches mentioned in this area brings up a fresh avenue for expanding the current research methods.*

Design/Methodology/Approach: *The study results that are provided in different scholarly works are acquired from secondary sources, such as scholarly journal publications. This review study analyses these articles and highlights their interpretations.*

Findings/Result: *This research focuses on several cataloguing methods for isolated identifying images and visions. When developing research topics in the idea of inaccessible detecting geographic information systems, the gaps discovered during analysis using various methodologies have made things simpler.*

Research limitations/implications: *This study examined a range of AI tool uses. The scope of this work is rivetted to a assessment of the many machine-learning implementation strategies for analysis and prediction. More research might be done on the many deep learning constructions for image and video classification.*

Originality/Value: *The articles chosen for this study's review are from academic journals and are cited by other authors in their works. The articles that were selected for the examination have a connection to the investigation and research plan described in the paper.*

Paper Type: *Literature review paper.*

Keywords: Machine learning, neural network, Semantic, visual description, image, vision.

1. INTRODUCTION :

Due to rapid improvements in the internet on computer vision several semantics contexts like Text, images and videos are can be seen nowadays where this made easily using some of the current systems like Deep learning technologies and neural networks, where explanation of the language can be made and several operations can be performed on images, text and videos where these are related with each other and classification and modification can be made through it and video summarization built on the language, qualities are done. The human behaviour or activity, biological images, and prophecies are finished using these techniques like vehicle disasters are prevented through it and use sensors and images from it and producing the high-quality image for the pattern recognition process. Thus, these semantics description operations are performed through it. Convolutional features that are effective are crucial for estimating saliency, yet learning effective saliency features is still difficult. Global semantic data from the top convolutional layer is conveyed to shallower layers via multi-path recurrent connections, which inherently improves the performance of the entire network [1]. Learning-based strategies are extremely advantageous for monocular depth estimates. To suggest an attention-driven loss for network supervision, use the long tail property and probe deeper into the far-off depth regions [2]. Using computer vision and machine learning approaches, automatic intrinsic memorability

prediction of images has just lately been studied. Semantic characteristics that represent scene elements, scene qualities, and evoked emotions [3]. By removing essential frames from the film, video summarising can be utilised to lessen this redundant information, making it easier to browse and index HD videos. An effective visual attention-driven approach for domain-specific DH video summaries. In order to extract critical frames, integral images are utilised to calculate multi-scale contrast, texture, curvature, and motion-based saliency features for each frame. These structures are combined by a linear weighted fusion approach to obtain a final saliency map [4]. By removing redundant information from videos through video summarising, it is possible to speed up the viewing and indexing of related videos [5]. The perception based on colour, texture, and motion is computed to realise the visual attention model [6]. Remote sensing (RS) picture scene classification has found deep learning (DL) based techniques to be popular. However, RS images generally contain many classes and can therefore be simultaneously linked with multi-labels. Utmost of the existing DL-based approaches accept that working out descriptions are annotated by single-labels [7]. Practitioners can better comprehend the behaviour and evolution of deep neural networks (DNNs) and get knowledge for architectural improvement by reading a thorough and understandable description of existing DNNs. Users of DNN Genealogy can gain knowledge about DNNs' architecture, performance, and evolutionary relationships [8]. The visual model is an executable data flow programme graph that was created automatically from language processing module data dependence declarations [9]. Deep learning techniques have delivered cutting-edge outcomes in several fields by using multiple processing layers to create hierarchical representations of the input. In the field of natural language processing, numerous model architectures and techniques have emerged (NLP) [10].

2. OBJECTIVES :

This review's main goal is to compile the findings of various studies in the field of image or video analysis, examine the methods used to make predictions and suggest ways to bridge knowledge gaps and pursue additional studies in this area.

(1) To analyse and implement, motion, global, and local entities-based feature extraction.
(2) To exploit local entities-based feature extraction
(3) To propose and implement the extract features using a range of deep neural trained networks.
(4) To implement an encoder-decoder based LSTM.
(5) To implement the attention mechanism in the encoder and decoder-based architecture.
(6) To experiment with various combinations of recent word embedding.
(7) To implement and analyse the framework on real-world video captioning datasets.
(8) To analyse the model's outcomes, performance indicators etc.

3. METHODOLOGY :

The paper consists of a deep learning algorithm for computer vision of interacting with the low-level image to high level image. The pre-processing stage focuses on preparing the images derived from input view and pre training the datasets for prediction and CNN, RNN, TNN features used for the feature extraction and prediction, video summarize by comparing with each video frame and language pattern and biological and pattern recognition using neural network system with the different architecture provided by it. This paper focus on the material that has been published in scholarly journals between 2004 and 2022. Research papers, review articles, and case studies are gathered from various national and international journals using the Google Scholar search engine. Books, journals, theses, and websites that have been published in the field of visual prediction are among the several alternative sources. The methods employed in visual analysis and prediction are examined using the ABCD analysis tool.

4. REVIEW OF LITERATURE /RELATED WORK :

There are numerous models and comparative studies of the various methods in the literature. For the purpose of presenting the research findings, related research publications are gathered and used. With the use of Google Scholar's Advanced Search feature, a number of publications dated between 2015 and 2022 that contain the keyword "images and visual prediction" were gathered. Table 1 lists the findings from this search.

Table 1: Lists research publications on visual analysis and semantic segmentation.

S. No.	Area of Research	Focus	Outcome of Research	Reference
1.	Graph neural networks in computer vision.	A Survey on Graph Neural Networks and Graph Transformers in Computer Vision: A Task - Oriented Perspective	In this survey on graph neural networks and graphs transformer in computer vision where it is a task-oriented perspective using the deep learning concepts with adding neural networks and CNN model where variety of vision problems are stored through it. then it has included image classification, object detection, semantic segmentation, take input modalities as humans do,2D,3D images and videos (natural and medical image) and multi model input that are image and text and then background and categorisation are made used of GNUS and graph transformers used in computer vision and several networks used for approximation like Graph convolution network(GCN),GAT(graph attention network).in this video understanding and action recognition are made ,vision and language understanding task performed and then visual questioning and answering(VQA) are performed in image. the graph representation of recurrent GNNs, convolutional GNN's, chebshey spectral CNN, Graph transformers for 3D data and scene graph generation (SGG) and other video segmentation are made in it.	Chaoqi Chen et al. (2022). [11]
2.	Analyse videos by using deep learning technology	Video Summarization Using Deep Neural Networks: A Survey	In this paper the video summarization using the deep-Neural network and survey by considering highest video uploaded platforms like youtube, video summarization is made and it includes extraction, performing, clustering based key selection using deep learning algorithms and create summarize by modelling human understanding preference using attention model, semantics visual content and statistical processing low level features of the video. Vectors are extracted from frame level and sample strategy and proposed the deep learning algorithm based on summation algorithm. deep learning approaches are made, general remark of deep leaning are approaches in video summarization and evaluation are made on video summarization using datasets, protocol and measures of its performance comparision of are made quantitatively with demos and future direction using deep learning algorithm based on video summarization.	Evlampios Apostolidis et al. (2021). [12]
3.	Analysing, grounding and synthesis	Trends in Text to Image	In this trend in Text to image (T2I) using the generative adversial network. The Generative adversial network(GANS) that is text to	Venkatesan, R. et al. (2021). [13]

	images by using NLP Technology	Synthesis (T2I) using Generative Adversarial Networks	image synthesis where successfully applying to natural language process(NLP) and computer vision and human like intelligence and brain visualize is used for analyze grounding and synthesis images and useful for the crime scene investigation. In this deep attention architecture used for enhancing image quality and reality,text and image alignment after synthesis is investigated and agreement are harkened by it.data sets like COCO,CVB,OXFORD are used .perform and evalution made to produce the high quality images and wide range of image in GANS synthesizing.	
4.	Text editing analysis using NLP	Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language	The paper consists of text adaptive generative adversal networks manipulating image with the natural language. In this photos based on smartphones clicks, photographs which images are edited using different software like or tools photoshops. So in this manipulation is made using NLP the images are manipulated by focuses on modifying visual attributes of on objects characterizing the texture, color of the object, in this they have proposed GAN structure where image to image domain translation and text to image synthesis methods and text adaptive generative adversal networks where description match the image. Several experiments are performed for the image and categories to it and quantitative results are made and analysis are done.	Seonghyeon Nam et al. (2018). [14]
5.	Examining image captioning architecture using Deep Learning	What is not where: the challenge of integrating spatial representations into deep learning architectures	In this that they examine the integrated visual and linguistic information where for this deep learning architecture is created, high level visual features like object parts from the labelled image, natural scene, where it is DL is used for image and language modelling task where standard DL image captioning architecture where image captioning used CNN process for inputs and encodes information from the image as vector, CNN(conventional neural network) where image recognition of hand writing digit recognition and local visual feature and combine to high order feature, RNN (recurrent neural networks language model) processing sequential data that is language based on its operation are performed by it. Grounding spatial language in perspective and spatial language in deep learning are performed for it.	John D. Kelleher et al. (2018). [15]
6.	Various image management by	Towards Accurate	It is based on the realisation of automated and intelligence transportation system are	Khorramshahi, P. (2021). [16]

	using Margin Removal Technology of NN.	Visual and Natural Language-Based Vehicle Retrieval Systems	made for the risk of accident happened to the vehicle. In deep neural networks quite possible of the traffic system management and speed and retrieve of vehicle different attributes and description, in this vehicle re-identification is made like model, colour and year where natural language based retrieval is made use of it. Margin removal (MR) where large margin can be viewed as background disasters and discrimination regions on primitive region. SCR (same camera removal) in this where the capture of the images of vehicle are identified by it and tracking of image is done.	
7.	CNN Module for pattern recognition.	Sentiment-Aware Deep Recommender System with Neural Attention Networks	In this, where the recommendation aims for today's products of information overhead thereby assists the customer to get their best choices from the various alternatives choices and based on the similar assumption habits in the part share similar items in future, filtering methods are made based on the matrix factorisation (MF). CNN module for pattern recognition are used in it. Architecture proposed in SDRA model for both semantics and conceptual information of words. Quantity analysis made on the attention visualization by identifying words using neural attention networks.	Da'u, A. et al. (2020). [17]
8.	Classification of images using CNN and RNN.	ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos	In this where machine learning is computer vision, used to interpret the real world through some from sensor it is used for learn from the images and videos by processing each frame and separate image. CNN used for image classification and RNN sequentially. Human activities recognition move complete task for that RNN are or general activity recognition model are used. Some operations performed on background of the objects like positional encoding sequence to sequence model, TNN, self-attention by neural networks and multi headed attention using TNN and several operations performed through it.	Wensel, J. et al. (2022). [18]
9.	Quality estimation of images and videos by applying learning algorithms	SABV-Depth: A biologically inspired deep learning network for monocular depth estimation	In this computer vision which enables real scene from the flatten image and monocular depth estimation to sense the high quality from the single image and prediction pixel value of the image and with classification into supervised learning , semi supervised and self -supervised where different prediction are made with the difficult type of image .several operation performed as bio network mapping ,visual information process and hierarchical phase in brain. Quantative	Wang, J. et al. (2023). [19]

			analysis made in the brain interaction mechanism. Training datasets are for training visualization for the biological different image training is made and biological related predictions are made with it.	
10.	systematic approach to evaluate video quality objectively	The Video Quality Experts Group objectively assessed video quality and related calibration methods (VQEG)	Only the NTIA Generic Model was among the top video quality estimators for both the 525-line and 625-line video tests. As a result, the NTIA General Model and its related calibration methods were accepted as a North American Standard by the American National Standards Institute (ANSI). In many applications, advanced video quality assessment (VQA) approaches strive to analyse the perceptual quality of videos, but they frequently lead to an increase in computational complexity.	Pinson, M. H. et al. (2004). [20]

To enable the user to interact with other people and things in the environment more precisely, every component of each video frame would be segmented in a human-computer interaction system. Deep semantic knowledge of the world and what objects are components of larger wholes are necessary for segmenting a picture. The summary of the literature search conducted using Google Scholar for publications published between 2010 and 2023 using the phrase "Semantic context in visual detection using deep learning" is shown in Table 2.

Table 2: Literature study on Semantic context in visual detection using deep learning.

S. No.	Area of Study	Focus	Outcome	Reference
1.	Semantic Segmentation	Segmentation is used to identify groups of pixels that fall into well-defined categories	A deep learning system called semantic segmentation assigns a label or category to each pixel in an image. It is used to identify groups of pixels that represent various categories. One of the major difficulties in the lengthy history of computer vision has been semantic segmentation, which is the capacity to divide an unknown image into several components and objects.	Guo, Y. et al. (2018) [21]
2.	Conventional techniques for object detection	By creating intricate ensembles that integrate numerous low-level visual elements, performance can easily plateau.	In order to solve the issues with traditional architectures, more effective tools that can learn semantic, high-level, deeper features are being offered as a result of deep learning's quick development. In terms of network architecture, training methodology, and optimization functionality, these models behave differently.	Zhao, Z. Q. et al. (2019). [22]
3.	A semantic segmentation's effectiveness	Research on semantic segmentation in the deep learning period: advances and problems.	Computer vision's difficult task of semantic segmentation. Deep learning approaches have significantly enhanced semantic segmentation performance in recent years. Several innovative techniques have been put forth. Semantic segmentation, a fundamental	Hao, S. et al. (2020). [23]

			yet difficult task in computer vision research, assigns a category name to each pixel of an image. Several real-world applications gain from this task because semantic segmentation is able to deliver the category information at the pixel level.	
4.	Using Deep Neural Networks for Photo Fixing	By artistically refining their images with aesthetic colour and tone alterations, photographers can evoke striking visual sensations.	An exciting alternative to manual labour is using an automated algorithm, however such an algorithm confronts several challenges. Several photographic techniques rely on minute alterations that are influenced by the semantics and even the content of the image. Deep learning has demonstrated exceptional aptitude at solving complex issues. This inspired everyone to investigate the application of deep neural networks (DNNs) to photo editing.	Yan, Z. et al. (2016). [24]
5.	Utilising deep neural networks for self-supervised image recognition	Significant advancements in deep learning have been made in image understanding tasks like object identification, picture categorization, and image segmentation. However the success of picture identification mostly depends on supervised learning, which necessitates a large number of labels that have been manually annotated.	Self-supervised learning is a type of unsupervised learning that enables the network to learn rich visual properties that aid in carrying out subsequent computer vision tasks like picture classification, object recognition, and image segmentation. In the field of computer vision, different self-supervised methods utilising pretext tasks have recently become more popular. These user-designed pretext activities aid in the learning of detailed representations from the input visuals. Extracting valuable data properties that support the construction of classifiers or other predictors is made simpler by learning representations of the input signal.	Ohri, K., and Kumar, M. (2021). [25]
6.	Visual classifications by using semantic segmentation	Image categorization, object identification, and border localisation are necessary for semantic segmentation.	Deep neural networks are highly good at semantic segmentation, which involves assigning a class of objects or non-objects to each region or pixel in an image. Semantic segmentation is crucial for image analysis tasks and plays a significant role in image comprehension. It can be used in a variety of artificial intelligence and computer vision applications. Object identification, border location, and	Lateef, F., and Ruichek, Y. (2019). [26]

			image classification are necessary for semantic segmentation.	
7.	Object detection, image captioning, event detection and recognition.	To solve sophisticated or high level vision tasks including segmentation, scene understanding, object tracking, image captioning, event detection, and activity recognition, object detection serves as the foundation for image understanding and computer vision.	Visual object recognition intends to locate objects of specific target classes inside a given image with high localization accuracy and to label each object instance with the appropriate class. Using properly thought-out feature descriptors to acquire embedding for a region of interest was the main emphasis of the majority of the effective classical approaches for object detection. The new deep learning-based algorithms significantly outperformed the old-school detection systems. A biologically inspired structure for computing hierarchical features is the deep convolutional neural network. In contrast to hand-crafted descriptors used in traditional detectors, deep convolutional neural networks generate hierarchical feature representations from raw pixels to high level semantic information, which is learned automatically from the training data and shows more discriminative expression capability in complex contexts.	Wu, X. et al. (2020) [27]
8.	High-Resolution Representation in Depth Learning to Recognize Images	For position-sensitive vision issues including human posture estimation, semantic segmentation, and object detection, high-resolution representations are crucial.	The low-to-high process tries to produce high-resolution representations, whereas the high-to-low process aims to produce low-resolution and high-level representations. A high-resolution network that produces precise and accurate key-point heat maps for human posture assessment.	Wang, J. et al. (2020) [28]

5. ANALYSING MOTION EXTRACTING FEATURES FROM GLOBAL AND LOCAL ENTITIES :

One of the most active areas of research in computer science and artificial intelligence is natural language processing (NLP), which uses software to process massive volumes of linguistic data. In this field, deep learning technologies have been developed and put to good use [29]. By choosing the most instructive segments of the video content, video summarising systems seek to produce a succinct and comprehensive overview [30]. The most advanced techniques are those that use contemporary deep neural network architectures. Several strategies have been developed during the past few decades [31]. Image captioning is a vital technique to explore the transfer of artificial intelligence visual perception to high-level semantic understanding since it combines the tasks of natural language processing with computer vision [32]. It is extensively utilised in a variety of industries, including image retrieval, blind navigation, early childhood education, human-computer interface, and safe aided driving. For text mining and information retrieval, the selection of text feature items is a fundamental and significant issue [33]. The foundation of many text processing techniques is text feature extraction, which extracts

text information to represent a text message [34]. The hidden layer of each machine serves as the visible layer for the following machine in a deep neural network (DNN), which may be thought of as a composite of straightforward, unsupervised models like restricted Boltzmann machines [35]. A current area of interest for the world of computer vision and video processing is human motion analysis. The numerous uses of this field, including monitoring and surveillance systems, serve as a driving force behind its research. The KNN, Neural network, SVM, and Bayes classifiers are trained using the collected features to recognise a collection of seven human activities [36].

6. LOCAL ENTITIES-BASED FEATURE EXTRACTION :

Convolutional Neural Networks (CNNs), a type of deep learning technique, have proven effective for various biometric systems and can automatically extract distinctive properties [37]. The final classification is carried out using a basic back propagation neural network using the features recovered by a transferred deep convolutional neural network, which is employed as a feature extractor [38]. Text document images contain both the visual signals and the associated text within the image, in contrast to the general image classification problem in the computer vision field. It is still difficult to combine these two disparate modalities and use textual and visual attributes to categorise text document photos. A cross-modal deep network that may be used to extract both the textual and visual content from document images. A picture of a document may be used to extract both the textual and visual content using a cross-modal deep network. To extract picture and text features, using NAS Net-Large and Bert [39]. Google first created Inception in 2014, and it was improved during the following two years [40]. An essential issue is the classification of entities using network structure. Practice-based networks are sparse, with lots of noisy and missing linkages. Intra-network classification has made use of statistical learning methods [41]. Information extraction (IE) may structure and semanticize unstructured multi-modal information and is a crucial link in the domains of natural language understanding and information retrieval [42]. Search engines are essential for retrieving information from web pages. The retrieval is based on content-based information extraction techniques or statistical searching approaches [43]. The majority of linguistic resource development research has concentrated on identifying three properties: subjectivity, orientation, and strength of term attitude. Support vector machines are regarded as a must-try in today's machine learning applications since they provide one of the most reliable and accurate techniques among all well-known algorithms [44]. In view of the diverse learning knowledge eco system of artificial intelligence of traditional cross-media retrieval technology, which is unable to obtain retrieval information timely and accurately, a study on cross-media intelligent perception and retrieval analysis technology based on deep learning education is being conducted [45]. In order to extract information from a given unstructured or semi-structured input source, representations of these elements must be found [46].

7. EXTRACTION FEATURE USING A RANGE OF DEEP NEURAL NETWORKS :

The widespread use of deep learning technology has increased as a result of the quick development of information and communication technologies [47]. A global gating unit for each pair of layers is used in the proposed RNN, called a gated-feedback RNN (GF-RNN), to enable and regulate signals flowing from upper recurrent layers to lower ones. This extends the present method of stacking multiple recurrent layers [48]. Combining standard audio features, like mel frequency cepstral coefficients, with visual features in video analysis often results in very slight gains. The neural network's output is a classification of the sentiment as either negative, positive, or neutral. Applications for recognising human activity in the physical world include intelligent video surveillance, customer qualities, and shopping behaviour analysis. Due to crowded backgrounds, occlusions, differing viewpoints, etc., accurately recognising activities is a very difficult process. The majority of current methods assume certain things about the setting in which the video [50]. The use of standard video for motion monitoring has the potential to be inexpensive and simple. Large publicly accessible datasets and contemporary computational techniques, such as deep learning, have made it possible for pose estimation algorithms, like Open-Pose, to generate estimates of body pose from common video under a variety of lighting, activity, age, skin colour, and angle-of-view conditions [51]. Alex Nets, GoogLeNet, and ResNet50 are the most widely used convolution neural networks for object detection and object category categorization from videos [52].

8. LSTM BASED ON ENCODER-DECODER :

The neural encoder-decoder framework has substantially improved machine translation's state-of-the-art. In recent years, many academics have used encoder-decoder based models to tackle challenging problems including text summarization, image/video captioning, and text/visual question answering [53]. The process of creating a written description that explains the objects and activities shown in a given photograph is known as image captioning [54]. It links the artificial intelligence disciplines of natural language processing and computer vision. Image understanding and language modelling are the respective topics of computer vision and natural language processing [55]. In a multi-step prediction, the encoder-decoder paradigm gives excellent robustness. Deep learning has received a lot of attention. When compared to conventional machine learning algorithms, deep learning algorithms can automatically extract features from raw data using their multi-layer networks, which reduces the manual work of feature engineering [56]. Despite the fact that the encoder-decoder modelling technique has been employed in a variety of applications throughout the literature, the framework is designed and interpreted from a control engineering standpoint. The approach's fundamental benefit is that the model(s) may be quickly built from a model definition, which mostly consists of a list of inputs and outputs. The model(s) might then be trained automatically using data [57]. In applications involving human movement, individual mobility prediction is essential. Individual mobility is modelled in studies of location prediction as a sequence of time-stamped locations, and the prediction problem is frequently framed as one of foretelling an individual's next location [58].

9. ENCODER AND DECODER BASED ARCHITECTURE ATTENTION MECHANISM :

It has been demonstrated that contextualized embedding like BERT and GPT significantly enhance NLP tasks. Recent attempts have been made to improve the skip-gram model by including syntactic context information using GCN [59]. Recent months have seen a significant improvement in performance for various natural language processing (NLP) applications using contextual word embedding models like ELM and BERT. The generation of contextual word embedding using word constituents as input has been proposed by a new wave of algorithms based on training language models, such as Open AI GPT and BERT. By combining the word pieces, these algorithms can produce representations for words that are not in the user's vocabulary. Recently, the state of the art for many NLP tasks has been steadily improved by fine-tuning pre-trained language models that have been trained on massive corpora [60]. Transfer learning, attention processes, and a variety of methods are combined in an encoder-decoder-based, rebuilt architecture for molecular picture captioning to increase effectiveness and adaptability in various datasets. The use of encoder-decoder-based DL in molecular image captioning is a result of its successful implementations in a variety of image and language tasks [61].

10. NLP METHODS BASED ON BERT FOR CLASSIFICATION :

The fact that certain sorts of semantic and syntactic links are captured by word embeddings is one of their most striking characteristics. Pre-trained language models, like BERT, have recently excelled in a variety of NLP tasks, achieving ground-breaking outcomes [62]. The NLP benchmarks have reached a new level attributable [63] to contextual language models (CLMs). Using word embeddings from CLM in downstream tasks like text categorization has become the norm [64]. To analyse and ascertain the attitudes expressed in various textual documents [65], such as social media posts, online product evaluations, and so forth, word embedding approaches have been developed in the literature [66].
The new dynamic, context-sensitive, token-based embedding from language models like BERT have, in the opinion of the NLP community [67], supplanted the more traditional static, type-based embedding like word2vec or fastText because of their superior performance [68]. Transformers' Bidirectional Encoder Representations BERT, among other pre-trained language models, significantly outperforms industry benchmarks in eleven NLP tasks, including sentence-level sentiment classification, setting a new standard for text representation [69]. To capture the semantics and context of words, BERT used the notion of contextualised word embedding [70]. BERT has demonstrated to be a straightforward yet effective language model that performed at a high level on eleven NLP tasks [71].

11. MOST MODERN WORD EMBEDDING NLP PERMUTATIONS :

The popularity of video sharing platforms has created new difficulties for video retrieval, such as the quick increase in video length and variety of material [72]. The most effective method for captioning videos has been developed using attention-based models that link key visual cues to video phrases. Existing studies often use equal interval frame sampling, frame-level appearance modelling, and motion modelling, which might result in redundant visual information, sensitivity to content noise, and needless computation costs [73].

The computer vision task of video captioning, which automatically uses natural language sentences with a grasp of the embedded semantics, is quite difficult. A technique for creating video captions using multimodal feature attention and discrete wavelet convolutional neural architecture [74].

Dense video captioning is a very difficult undertaking since it calls for a thorough understanding of the contents of the video as well as contextual reasoning of specific occurrences in order to accurately and faithfully describe events in a video [75]

Robots are typically outfitted with cameras to investigate indoor environments, and it is anticipated that the robot will be able to accurately describe the environment using natural language. Although there has been significant progress in picture and video captioning technologies, particularly with regard to several publicly available datasets, the caption produced from video of inside scenes is still not insightful and cogent enough [76].

Identifying the objects seen and speculating on their relationships are necessary for describing an image's main event. The relationships between the items in a picture are captured via visual dependency representations, and it is hypothesised that these representations can enhance image description. The models outperform methods that depend on object proximity or corpus knowledge to generate descriptions in an image description challenge on both automatic metrics and subjective evaluations [77].

Sensorimotor (movement, grabbing), interactive (joint manipulation of an object), language, and autonomous self-exploration or exploration under the supervision of a human teacher are the four categories into which robot talents can be categorised [78]. Robot learning can be closely related to topics like adaptive control, which uses dynamically adapting controllers to improve sensorimotor skills; reinforcement learning, which helps with understanding, acting, and planning; and developmental robotics, which uses higher levels of autonomous learning modalities similar to those found in young children, where lifelong learning is expected to be cumulative and progressively more complex.

The research using the MS-COCO dataset shows how NLP methods can be used to predict visual descriptions from images [79]. This work has many real-world applications, including helping the blind access visual content and automatically creating captions for pictures on social media platforms. There is now a wealth of literature on methods for drawing reliable conclusions from conventional (non-text) datasets but applying these methods to natural language data presents fresh, fundamental difficulties [80].

12. ANALYSIS OF REAL-WORLD VIDEO CAPTIONING :

Modern image captioning models have recently surpassed human performance based on the most widely used metrics, including BLEU, METEOR, ROUGE, and CIDEr. There are many alternative captions that express various concepts and degrees of detail that might be attractive to different humans because an image has many concepts and levels of detail. [81]. Few scholars have examined the impact of new technology, such as video captioning, on language acquisition, particularly on listening, despite the importance of technology in assisting learning [82].

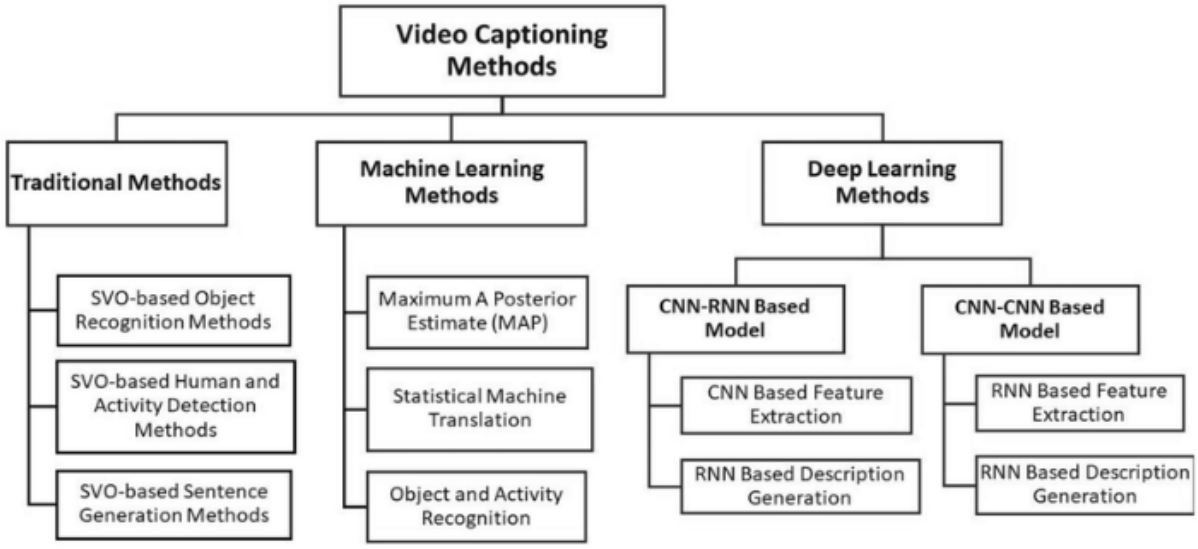

Figure 1: Different methods used for Video Captioning

Fig. 1: describes various methods of video captioning [83]

The automatic creation of words in natural language that summarise the information in a particular movie is known as video description. It can be used for video subtitling, human-robot interaction, and assisting the blind. It is difficult to determine the contributions to the correctness or errors of the visual features and the chosen language model in the final description, making analysis of video description models complex [84].

13. THE USE OF NATURAL LANGUAGE PROCESSING TO FORECAST VISUAL DESCRIPTIONS :

Using Language and Vision Together to Produce Natural Language Descriptions of Wild Videos: The emphasis is on choosing content to create phrases that will describe videos. Due to the abundance of video actions and objects and the dearth of training data, a graphical model for fusing noisy computer vision detections with statistical linguistic knowledge extracted from vast text corpora has been developed [85]. Natural Language Processing and Computer Vision: New Directions in Multimedia and Robotics, there are numerous ways to share meaning in human communication, including writings, gestures, sign languages, and facial expressions. Language without perception becomes an abstraction; on the other hand, perception without language would be limited to simple conditioned responses [86]. The challenge of creating a stochastic model to forecast the behaviour of a random process is dealt with through statistical modelling. Speculators on Wall Street create models based on historical stock price movements to forecast future fluctuations and change their portfolios to profit from the foreseen future [87]. The state of the art for many different natural language processing tasks has significantly improved thanks to language model pre-training [88]. Pre-trained LMs can be adjusted to adapt to subsequent tasks. They learn contextualised text representations by predicting words based on their context utilising a huge amount of text data. Using artificial intelligence to process, arrange, and extract embedded information from texts, or natural language processing (NLP), a branch of linguistics and computer science [89]. Any statement that disparages an individual or a group on the basis of a trait such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or another characteristic is usually referred to as hate speech [90]. Basic word filters do not adequately address this issue, necessitating the use of natural language processing. The task of automatically creating natural language image descriptions at the sentence level has drawn more attention in recent years. Visual recognizers have previously discovered object instances in advance, and they focus on a particular aspect of description production [91].

14. NATURAL LANGUAGE PROCESSING FOR PREDICTING VISUAL DESCRIPTION :

To aid in the description of images and videos, linguistic and visual data have been combined in a number of recent initiatives. The prep dependence is used to determine object-place relationships,

ensuring that the noun changed by the preposition is one of our recognisable places [92]. A picture is frequently represented in existing work as a collection of image areas. It is challenging to explain what is going on since bags of regions only record which items co-occur in an image; they cannot communicate how the areas connect to one another [93]. The most common method for creating the descriptions entails finding instances of pre-defined concepts in the image, then using reasoning to create image descriptions [94]. These efforts can be unified, given a single nomenclature, assisted in identifying underlying causes, and allowed to coordinate countermeasures by using a common definition and framework of predictive bias. Based on a thorough review of the pertinent NLP literature, with input from a few social science and related publications [95]. Using a human-and-model-in-the-loop training technique to compile a fresh standard for interpreting natural language [96]. The difficult task of computer vision and natural language processing is image captioning. The main difficulty is to use machines to extract semantic information from photos and translate it into human language [97]. Deep learning methods are adept at handling the difficulties of image captioning. The majority of picture captioning models use an encoder-decoder design, with the encoder receiving input in the form of abstract image feature vectors. One of the most effective techniques makes use of feature vectors that are derived from the region suggestions that an object detector provides [98]. Although there are many different image caption systems, image caption can be used for image retrieval, video caption, and video movement [99]. The discipline of Natural Language Processing has seen considerable success using deep learning techniques. Bidirectional Encoder Representations from Transformers (BERT), an NLP method, was recently created at Google to enhance language understanding challenges [100]. Approaches based on natural language processing have recently attracted interest for the legal systems of various nations.

15. RESEARCH GAP :

The examination of the literature demonstrates that different Machine Learning methods are used to classify images or visions. Quantifying the changes in video is done afterwards using change detection algorithms.
Research Gap 1: The usability and accuracy of mapping systems have scope for improvement.
Research Gap 2: Calculations can be made to get the ideal index for the chosen region of interest.
Research Gap 3: Hyper factors can be used to adjust classification models.

16. RESEARCH AGENDA BASED ON RESEARCH GAP :

(1) How to perform categorising and comparing the semantic contents?
(2) What are the pre-processing methods suitable for visual semantic?
(3) Which machine learning algorithm gives optimal results in developing an image and video classifications?
(4) How to improve the classification accuracy?
(5) How to generate a visual and semantic context?
(6) How to perform visual and quantitative change detection analysis?

17. ANALYSIS OF RESEARCH AGENDA :

By analysing scientific and applied research work on mapping the study region, a technical system for the development of vision-based images and technology can be built. The initial collection of spatial data, software selection, preprocessing, creation of layers into themes, conditional character processing, printing, and other steps involved in creating, captioning visions are all part of this technical system.

18. FINAL RESEARCH PROPOSAL/PROBLEM IN CHOSEN TOPIC :

(1) Retrieving, preprocessing, and analysing data collected from various websites.
(2) Mapping the identified inputs according to the study purpose.
(3) To use object-based change detection technique and track the differences in video captions before and after updating.

19. ABCD RESEARCH PROPOSAL ANALYSIS :

The research process calls for a set of phases to be followed and completed methodically. The DDLR model discussed in this article focuses [101] on the flow of the research process and aids in the creation

of a dependable and solid research methodology. A qualitative analysis known as ABCD analysis is provided and contrasted with other approaches like SWOC, CPM analysis [102], etc. within the framework for studying and analysing corporate strategies and models. The writers of this article [103] have discussed how ABCD analysis can be used in company case analysis approaches as a research methodology. SWOC analysis is used to determine the benefits, drawbacks, opportunities, and difficulties of a product. The writers of SWOC analysis [104] for higher education institutions have reviewed and used it. ABCD analysis is also applicable for determining the potential of the research methodology designed to carry out research in a particular area.

Advantages:

(i) The advancements in video and image captioning allow for practical applications such as automatic video subtitling, surveillance footage, text-based video retrieval that is affordable for blind users, video understanding, and multimedia recommendation.

(ii) Helping those who have varying degrees of vision impairment is one of them, as are self-driving cars, sign interpretation, human-robot interaction, and intelligent video subtitling.

Benefits:

(i) Real-world applications like automatic video subtitling, surveillance footage, text-based video retrieval afordability for blind users, video comprehension, multimedia recommendation is made possible by video and image captioning advances.

Constraints:

(i) Assessment of classification accuracy

(ii) Dependability of data

(iii) Lacking historical vision truth data

Disadvantages:

(i) It would remain an unanswered question how to accurately identify the right items in pictures, even if they are cloudy or occluded.

(ii) The visual encoder transmits video and linguistic data unidirectionally to another level.

20. SUGGESTIONS FOR IMPLEMENTING THE PROPOSED RESEARCH ACTIVITIES :

The different change detection algorithms that were discovered in this study will first be used to the remotely captioned photos and videos that were gathered for the study. To obtain reliable results, the method will be identified and extended with finetuning based on how well the approach fits the chosen region of interest. The photographs that would be gathered through a website would be used for analysis.

21. LIMITATIONS OF THE PROPOSAL :

The research proposal presented in this study provides a general overview of how video and picture data be used for captioning and prediction in order to perform future and identify change. This proposal does not specifically describe the research site or the type of technology that will be employed to track change.

22. CONCLUSION :

The framework fully explores the context like videos, text and images where deep learning techniques, CNN (conventional neural network), RNN techniques are used for the description of the object. where CNN and deep learning features which would able to go through deeply with images and videos summarization for better results or prediction can be made. Also used in the biological and spatial related images prediction based on the classification and modification are made and GAN features are used for description of the language's techniques by it and varieties of evaluation metrics.

REFERENCES :

[1] Zhang, X., Wang, T., Qi, J., Lu, H., & Wang, G. (2018). Progressive attention guided recurrent network for salient object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition, 1*(1), 714-722. Google Scholar ↗

[2] Jiao, J., Cao, Y., Song, Y., & Lau, R. (2018). Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. *In Proceedings of the European conference on computer vision (ECCV), 16*(1), 53-69. Google Scholar ↗

[3] Celikkale, B., Erdem, A., & Erdem, E. (2015). Predicting memorability of images using attention-driven spatial pooling and image semantics. *Image and vision Computing*, *42*(1), 35-46. Google Scholar↗

[4] Muhammad, K., Sajjad, M., Lee, M. Y., & Baik, S. W. (2017). Efficient visual attention driven framework for key frames extraction from hysteroscopy videos. *Biomedical Signal Processing and Control*, *33*(1), 161-168. Google Scholar↗

[5] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., & de Albuquerque, V. H. C. (2019). Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. *IEEE Internet of Things Journal*, *7*(5), 4455-4463. Google Scholar↗

[6] Ejaz, N., Mehmood, I., & Baik, S. W. (2013). MRT letter: Visual attention driven framework for hysteroscopy video abstraction. *Microscopy research and technique*, *76*(6), 559-563. Google Scholar↗

[7] Sumbul, G., & DemIr, B. (2020). A deep multi-attention driven approach for multi-label remote sensing image classification. *IEEE Access*, *8*(1), 95934-95946. Google Scholar↗

[8] Wang, Q., Yuan, J., Chen, S., Su, H., Qu, H., & Liu, S. (2019). Visual genealogy of deep neural networks. *IEEE transactions on visualization and computer graphics*, *26*(11), 3340-3352. Google Scholar↗

[9] Gaizauskas, R., Rodgers, P. J., & Humphreys, K. (2001). Visual tools for natural language processing. *Journal of Visual Languages & Computing*, *12(*4), 375-412. Google Scholar↗

[10] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75. Google Scholar↗

[11] Bhadani, R., Chen, Z., & An, L. (2023). Attention-Based Graph Neural Network for Label Propagation in Single-Cell Omics. *Genes*, *14*(2), 506-515. Google Scholar↗

[12] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, *109*(11), 1838-1863. Google Scholar↗

[13] Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)*, *55*(2), 1-25. Google Scholar↗

[14] Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *Neural Networks*, *144*(1), 187-209. Google Scholar↗

[15] Dobnik, S., Ghanimifard, M., & Kelleher, J. (2018, June). Exploring the functional and geometric bias of spatial relations using neural language models. *In Proceedings of the First International Workshop on Spatial Language Understanding*, *14*(1), 1-11. Google Scholar↗

[16] Khorramshahi, P., Rambhatla, S. S., & Chellappa, R. (2021). Towards accurate visual and natural language-based vehicle retrieval systems. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *21*(1), 4183-4192. Google Scholar↗

[17] Da'u, A., & Salim, N. (2020). Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, *53*(4), 2709-2748. Google Scholar↗

[18] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87-110. Google Scholar↗

[19] Wang, J., Chen, Y., Dong, Z., Gao, M., Lin, H., & Miao, Q. (2023). SABV-Depth: A biologically inspired deep learning network for monocular depth estimation. *Knowledge-Based Systems*, *263*(1), 110301-110309. Google Scholar↗

[20] Pinson, M. H., & Wolf, S. (2004). A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, *50*(3), 312-322. Google Scholar↗

[21] Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, *7*(1), 87-93. Google Scholar↗

[22] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, *30*(11), 3212-3232. Google Scholar↗

[23] Hao, S., Zhou, Y., & Guo, Y. (2020). A brief survey on semantic segmentation with deep learning. *Neurocomputing*, *406*(1), 302-321. Google Scholar↗

[24] Yan, Z., Zhang, H., Wang, B., Paris, S., & Yu, Y. (2016). Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, *35*(2), 1-15. Google Scholar↗

[25] Ohri, K., & Kumar, M. (2021). Review on self-supervised image recognition using deep neural networks. *Knowledge-Based Systems*, *224*(1), 107090-107098. Google Scholar↗

[26] Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, *338*(1), 321-348. Google Scholar↗

[27] Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. *Neurocomputing*, *396*(1), 39-64. Google Scholar↗

[28] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349-3364. Google Scholar↗

[29] Yang, H., Luo, L., Chueng, L. P., Ling, D., & Chin, F. (2019). Deep learning and its applications to natural language processing. *Deep learning: Fundamentals, theory and applications*, *2*(2), 89-109. Google Scholar↗

[30] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, *109*(11), 1838-1863. Google Scholar↗

[31] Yang, F., Su, X., Ren, J., Ma, X., & Han, Y. (2022, May). A Survey of Image Captioning Algorithms Based on Deep Learning. *International Conference on Image Processing and Media Computing (ICIPMC)*, *1*(1), 108-114. Google Scholar↗

[32] Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics*, *19*(6), 455-501. Google Scholar↗

[33] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, *37*(1), 98-125. Google Scholar↗

[34] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*(1), 42-49. Google Scholar↗

[35] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, *65*(1), 3-14. Google Scholar↗

[36] Babu, R. V., & Ramakrishnan, K. R. (2004). Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision computing*, *22*(8), 597-607. Google Scholar↗

[37] Labati, R. D., Muñoz, E., Piuri, V., Sassi, R., & Scotti, F. (2019). Deep-ECG: Convolutional neural networks for ECG biometric recognition. *Pattern Recognition Letters*, *126*(1), 78-85. Google Scholar↗

[38] Isin, A., & Ozdalili, S. (2017). Cardiac arrhythmia detection using deep learning. *Procedia computer science*, *120*(1), 268-275. Google Scholar↗

[39] Bakkali, S., Ming, Z., Coustaty, M., & Rusiñol, M. (2020, October). Cross-modal deep networks for document image classification. *IEEE International Conference on Image Processing (ICIP)*, *1*(1), 2556-2560. Google Scholar↗

[40] McNeely-White, D., Beveridge, J. R., & Draper, B. A. (2020). Inception and ResNet features are (almost) equivalent. *Cognitive Systems Research*, *59*(1), 312-318. Google Scholar↗

[41] Nandanwar, S., & Murty, M. N. (2016, August). Structural neighborhood based classification of nodes in a network. In Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, *22(*1), 1085-1094. Google Scholar↗

[42] Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. (2022). A Survey of Information Extraction Based on Deep Learning. *Applied Sciences*, *12*(19), 9691-9697. Google Scholar↗

[43] Vijayarajan, V., Dinakaran, M., Tejaswin, P., & Lohani, M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-centric Computing and Information Sciences*, *6*(1), 1-30. Google Scholar↗

[44] Vinchurkar, S. V., & Nirkhi, S. M. (2012). Feature extraction of product from customer feedback through blog. *International Journal of Emerging Technology and Advanced Engineering*, *2*(1), 2250-2459. Google Scholar↗

[45] Liu, H., & Ko, Y. C. (2021). Cross-media intelligent perception and retrieval analysis application technology based on deep learning education. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(15), 2152023-2152027. Google Scholar↗

[46] Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web*, *11*(2), 255-335. Google Scholar↗

[47] Kuo, R. J., & Kunarsito, D. A. (2022). Residual stacked gated recurrent unit with encoder–decoder architecture and an attention mechanism for temporal traffic prediction. *Soft Computing*, *26*(17), 8617-8633. Google Scholar↗

[48] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015, June). Gated feedback recurrent neural networks. *In International conference on machine learning*, *37*(1), 2067-2075. Google Scholar↗

[49] Takahashi, N., Gygli, M., & Van Gool, L. (2017). Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, *20*(3), 513-524. Google Scholar ↗

[50] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 221-231. Google Scholar↗

[51] Kidzinski, L., Yang, B., Hicks, J. L., Rajagopal, A., Delp, S. L., & Schwartz, M. H. (2020). Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, *11*(1), 4054. Google Scholar↗

[52] Sharma, N., Jain, V., & Mishra, A. (2018). An analysis of convolutional neural networks for image classification. *Procedia computer science*, *132*(1), 377-384. Google Scholar↗

[53] Asadi, A., & Safabakhsh, R. (2020). The encoder-decoder framework and its applications. *Deep learning: Concepts and architectures*, *866*(1), 133-167. Google Scholar↗

[54] Lyu, P., Chen, N., Mao, S., & Li, M. (2020). LSTM based encoder-decoder for short-term predictions of gas concentration using multi-sensor fusion. *Process Safety and Environmental Protection*, *137*(1), 93-105. Google Scholar↗

[55] Mishra, S. K., Rai, G., Saha, S., & Bhattacharyya, P. (2021). Efficient channel attention based encoder–decoder approach for image captioning in hindi. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(3), 1-17. Google Scholar↗

[56] Lapeyrolerie, M., & Boettiger, C. (2023). Limits to ecological forecasting: Estimating uncertainty for critical transitions with deep learning. *Methods in Ecology and Evolution*, *14*(3), 785-798. Google Scholar↗

[57] Ellis, M. J., & Chinde, V. (2020). An encoder–decoder LSTM-based EMPC framework applied to a building HVAC system. *Chemical Engineering Research and Design*, *160*(1), 508-520. Google Scholar↗

[58] Li, F., Gui, Z., Zhang, Z., Peng, D., Tian, S., Yuan, K., ... & Lei, Y. (2020). A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction. *Neurocomputing*, *403*, 153-166. Google Scholar↗

[59] Wang, Y., Cui, L., & Zhang, Y. (2021). Improving skip-gram embeddings using BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*(1), 1318-1328. Google Scholar↗

[60] Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, *470*(1), 443-456. Google Scholar↗

[61] Yi, J., Wu, C., Zhang, X., Xiao, X., Qiu, Y., Zhao, W., ... & Cao, D. (2022). MICER: a pre-trained encoder–decoder architecture for molecular image captioning. *Bioinformatics*, *38*(19), 4562-4572. Google Scholar↗

[62] Lim, S., Prade, H., & Richard, G. (2021). Classifying and completing word analogies by machine learning. *International Journal of Approximate Reasoning*, *132*(1), 1-25. Google Scholar↗

[63] Hu, Z., Cui, J., Wang, W. H., Lu, F., & Wang, B. (2022, April). Video Content Classification Using Time-Sync Comments and Titles. In 2022 *7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, *7*(1), 252-258. Google Scholar ↗

[64] Bhardwaj, R., Majumder, N., & Poria, S. (2021). Investigating gender bias in bert. *Cognitive Computation*, *13*(4), 1008-1018. Google Scholar↗

[65] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113. Google Scholar↗

[66] Subba, B., & Kumari, S. (2022). A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings. *Computational Intelligence*, *38*(2), 530-559. Google Scholar↗

[67] Vieira, V., Tedesco, P., & Salgado, A. C. (2011). Designing context-sensitive systems: An integrated approach. *Expert Systems with Applications*, *38*(2), 1119-1138. Google Scholar↗

[68] Ehrmanntraut, A., Hagen, T., Konle, L., & Jannidis, F. (2021). Type-and Token-based Word Embeddings in the Digital Humanities. In *CHR*, *2989*(1), 16-38. Google Scholar↗

[69] Chen, X., Cong, P., & Lv, S. (2022). A long-text classification method of Chinese news based on BERT and CNN. *IEEE Access*, *10*(1), 34046-34057. Google Scholar↗

[70] Shah, S. M. A., Taju, S. W., Ho, Q. T., & Ou, Y. Y. (2021). GT-Finder: Classify the family of glucose transporters with pre-trained BERT language models. *Computers in biology and medicine*, *131*(1), 104259. Google Scholar↗

[71] Yu, S., Su, J., & Luo, D. (2019). Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, *7*(1), 176600-176612. Google Scholar↗

[72] Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., & Lin, D. (2018). Find and focus: Retrieve and localize video events with natural language queries. *In Proceedings of the European Conference on Computer Vision (ECCV)*, *1*(1), 200-216. Google Scholar↗

[73] Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, *52*(6), 1-37. Google Scholar↗

[74] Bhooshan, R. S., & Suresh, K. (2022). A multimodal framework for video caption generation. *IEEE Access*, *10*(1), 92166-92176. Google Scholar↗

[75] Suin, M., & Rajagopalan, A. N. (2020, April). An efficient framework for dense video captioning. *In Proceedings of the AAAI Conference on Artificial Intelligence*,*34*(07), 12039-12046. Google Scholar↗

[76] Li, X., Guo, D., Liu, H., & Sun, F. (2021, May). Robotic indoor scene captioning from streaming video. *In 2021 IEEE International Conference on Robotics and Automation (ICRA)*,*1*(1), 6109-6115. IEEE. Google Scholar↗

[77] Elliott, D., & Keller, F. (2013, October). Image description using visual dependency representations. *In Proceedings of the 2013 conference on empirical methods in natural language processing*, *18*(21), 1292-1302. Google Scholar↗

[78] Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2016). Computer vision and natural language processing: recent approaches in multimedia and robotics. *ACM Computing Surveys (CSUR)*, *49*(4), 1-44. Google Scholar↗

[79] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. *In International conference on machine learning*, *37*(1), 2048-2057. Google Scholar↗

[80] Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., & Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, *10*(1), 1138-1158. Google Scholar↗

[81] Wang, Q., & Chan, A. B. (2019). Describing like humans: on diversity in image captioning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *1*(1), 4195-4203. Google Scholar↗

[82] Gowhary, H., Pourhalashi, Z., Jamalinesari, A., & Azizifar, A. (2015). Investigating the effect of video captioning on Iranian EFL learners' listening comprehension. *Procedia-Social and Behavioral Sciences*, *192*(1), 205-212. Google Scholar↗

[83] How different video captioning methods are used retrieved from https://www.researchgate.net/figure/Different-methods-used-for-Video-Captioning_fig1_349665373 Accessed on 02/03/2023. Google Scholar↗

[84] Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., & Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, *52*(6), 1-37. Google Scholar↗

[85] Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., & Mooney, R. (2014, August). Integrating language and vision to generate natural language descriptions of videos in the wild. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, *1*(1), 1218-1227. Google Scholar↗

[86] Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2016). Computer vision and natural language processing: recent approaches in multimedia and robotics. *ACM Computing Surveys (CSUR)*, *49*(4), 1-44. Google Scholar↗

[87] Berger, A., Della Pietra, S. A., & Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, *22*(1), 39-71. Google Scholar↗

[88] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, *54*(3), 1-40. Google Scholar↗

[89] Chan, C. R., Pethe, C., & Skiena, S. (2021). Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights*, *16*(1), 276-231. Google Scholar↗

[90] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. *In Proceedings of the fifth international workshop on natural language processing for social media*, *11*(1), 1-10. Google Scholar↗

[91] Dong, J., Li, X., & Snoek, C. G. (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, *20*(12), 3377-3388. Google Scholar↗

[92] Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65-170. Google Scholar↗

[93] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423-443. Google Scholar↗

[94] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., ... & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, *35*(12), 2891-2903. Google Scholar↗

[95] Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*, *1*(1), 1-9. Google Scholar↗

[96] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35. Google Scholar↗

[97] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, *51*(6), 1-36. Google Scholar↗

[98] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 539-559. Google Scholar↗

[99] Ding, S., Qu, S., Xi, Y., & Wan, S. (2020). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, *398*(1), 520-530. Google Scholar↗

[100] Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45-53. Google Scholar↗

[101] HR, G., & Aithal, P. S. (2022). The DDLR Model of Research Process for Designing Robust and Realizable Research Methodology During Ph. D. Program in India. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *7*(2), 400-417. Google Scholar↗

[102] Aithal, P. S. (2016). Study on ABCD analysis technique for business models, business strategies, operating concepts & business systems. *International Journal in Management and Social Science*, *4*(1), 95-115. Google Scholar↗

[103] Aithal, P. S. (2017). ABCD Analysis as Research Methodology in Company Case Studies. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *2*(2), 40-54. Google Scholar↗

[104] Aithal, P. S., & Kumar, P. M. (2015). Applying SWOC analysis to an institution of higher education. *International Journal of Management, IT and Engineering*, *5*(7), 231-247. Google Scholar↗

****