**Chapter 01: INTRODUCTION**

* 1. **Introduction**

One of the most creative and human-centered problems in the field of computer science in the era of digital transformation and artificial intelligence is the creation of full-length stories from images, a task that transcends simple description and delves into imagination, narrative logic, and contextual reasoning. The goal of this large-scale project, "Image to Story Generation," is to develop a deep learning-based system that can interpret a given image and produce a rich, coherent, multi-sentence story that captures not only the visual details but also the emotions, relationships, and implied dynamics within the scene.

Fundamentally, narrative is a part of human communication. It is how we express our feelings, preserve history, and share experiences. As the saying goes, "one image is worth a thousand words." The difficulty is in making it possible for a machine to convert those "thousand words" into a coherent story. Story generation calls for the model to go one step farther than typical picture captioning, which creates a succinct and frequently factual statement summarizing an image. It must synthesis the visual content into a temporal and emotional sequence that reads like a narrative authored by a human. As a result, the work is both semantically complicated and computationally demanding.

The project makes use of cutting-edge methods from two important artificial intelligence pillars—computer vision (CV) and natural language processing (NLP)—to address this issue. Convolutional Neural Networks (CNNs) are used to process the image in order to extract high-level features that reflect the different aspects in the image, including objects, spatial connections, and scene context. After that, these attributes are fed into a sequential model, usually a Transformer-based language model or an LSTM (Long Short-Term Memory) network, which creates the narrative word by word or phrase by sentence. The end effect is a coherent story that mirrors the image's content while incorporating a layer of inferred knowledge, like forecasting feelings, behaviors, or outcomes that might not be immediately apparent.

These datasets assist the model understand the relationships between visual aspects and narrative components, including grammar, tone, and flow. A combination of architectural design, attention mechanisms, and optimization techniques are used to address a number of challenges, including preserving coherence, avoiding repetition, guaranteeing diversity in storylines, and capturing the abstract elements of a story, such as suspense, mood, or character development.

Image-to-story generation has several significant real-world applications. By creating prompts from photos, it can be utilized in the educational field to educate creative writing. Through generating audio narrative, it can aid visually challenged people in understanding visual content. It can automate narrative for marketing images or personal photos in social media and content production. Furthermore, this technology has the potential to foster innovation and human-computer collaboration in domains such as digital art, journalism, entertainment, and psychology.

Additionally, this initiative addresses more general philosophical and technological ramifications: Are machines able to tell stories as well as people? Is it imagination, emotion, structure, or grammar that makes a tale "good"? By pushing artificial intelligence's limits in the direction of artistic expression, "Image to Story Generation" transcends its technological difficulties and becomes a profound investigation of what it means to understand and convey   
  
In conclusion, "Image to Story Generation" is a blend of deep learning, language generation, and cognitive modeling. It mirrors the greater transition in AI from strictly analytical systems toward models that can sense, conceive, and narrate. This project brings us one step closer to creating machines that can do more than just see; they can also monitor, understand,

In order to achieve this, the research combines sophisticated language models like Transformers or LSTMs for sequence generation with Convolutional Neural Networks (CNNs) for visual feature extraction. The model can learn both the visual semantics and the narrative structures because it is trained on datasets that include paired images and stories with many sentences. It is anticipated that after extensive training, testing, and refinement, the system will generate stories that are not only grammatically accurate but also creative and captivating, as well as generalize well to unseen images.  
  
The applications of such a system are far-reaching — from helping visually impaired individuals understand visual content to enhancing digital storytelling in entertainment, journalism, education, and social media.

Identifying objects, seeing scenes and events, figuring out linkages, and then turning this visual context into a narrative that resembles human storytelling are all steps in the process of creating a tale from a picture. A greater degree of abstraction is required for tale development than for conventional image captioning, which usually produces a straightforward descriptive statement. The machine must combine several incidents or feelings into a coherent and expressive story that not only explains what is observed but also provides context, depth, and potential future developments.

* 1. **Problem Statement**

The need for more in-depth, meaningful engagements with visual content is rising quickly in the modern world, where digital imagery is crucial for communication and information sharing. Computers can now comprehend visual data at a basic level thanks to traditional image processing tasks like object detection, scene classification, and image captioning. They are inadequate, nevertheless, in producing human-like narratives from images, which call for a greater degree of abstraction, emotional intelligence, and contextual reasoning. Machines, in contrast to humans, are not naturally able to deduce backstory, envision potential situations, or convey inventiveness from a single image.

The goal of this research is to close this gap by creating a deep learning-based system called "Image to Story Generation" that uses computer vision and natural language processing to create engaging, context-aware narratives from photos. The difficulty is not only in explaining what is observed, but also in explaining what might be occurring and why, which is still a difficult and unsolved topic in the field of artificial intelligence research.

**1.2.1 Limited Visual Data Interpretation:**

Current AI systems' limited comprehension of visual input is one of their biggest drawbacks. The majority of vision models are taught to do tasks like classifying scenes, identifying boundaries, and labeling objects, but they frequently fall short of understanding the image's underlying context. An photograph of a person kneeling by a headstone in a cemetery, for instance, might be classified merely as "person," "grass," and "stone," ignoring the image's potential narrative and underlying emotional significance. On the other side, because of their natural empathy and past experiences, humans are able to quickly understand the situational context, potential relationships, and emotional undertone.

The use of AI systems in domains such as journalism, narrative, education, and digital support is limited by their incapacity to go beyond surface-level recognition. The system's interpretation stays flat and robotic if it is unable to comprehend the relationships between entities, forecast behaviors, or infer emotions. More advanced models that can comprehend images on multiple levels, including spatial relationships, face expressions, and implied motion, are required to overcome this issue. Only then will we be able to develop narrative structures that reflect the depth and diversity of human vision and imagination.

**1.2.2 Absence of Creative and Contextual Language Production**

Even while natural language generation has advanced significantly, especially in captioning systems that explain what is in an image, the generated phrase's inventiveness and contextual nuance are still somewhat limited. The phrases produced by traditional captioning algorithms, such as "A man riding a horse" or "Two children playing in the park," are instructive but devoid of passion, plot, or interaction. Typically, these captions are constructed using basic sentence templates and taken from a predefined vocabulary. However, presenting facts is only one aspect of narrative. In order for a tale to be interesting and relatable, the system must be able to visualize a series of events, deduce motivations, incorporate dialogue or thoughts, and depict emotions.

The system's usability in practical applications like entertainment, education, or accessibility aids is limited by its incapacity to produce such imaginative, context-sensitive language. For example, visually impaired users may rely on audio narrations to appreciate and emotionally connect with what is happening in an image, in addition to simply knowing what is there. A simple caption isn't enough for this. In order to overcome these obstacles, the architecture must be able to connect visual aspects with narrative components like time, emotion, and interaction, and the system must be trained on a variety of expressive language data. One of the fundamental research issues in AI is still how to generate texts with this degree of inventiveness.

**1.2.3 Linking Language and Vision Models**

The successful integration of two essentially distinct domains—vision and language—is one of the primary obstacles in creating narratives from images. Language is sequential and symbolic, whereas visual data is spatial and continuous. Convolutional Neural Networks (CNNs) and other deep learning models for images are experts at recognizing edges, forms, textures, and objects by extracting patterns and characteristics from pixel input. However, language models, particularly transformer-based systems, are adept at comprehending the syntax, context, and word relationships found in textual data. These two models must be combined in a way that enables the system to convert abstract visual representations into coherent and meaningful language in order to create stories from images.

This integration is not simple. In order for the language model to generate coherent tales, high-dimensional picture embeddings must be mapped into a space that it can comprehend. Aligning visual attention with linguistic focus, or making sure that the model is "looking at" the same thing in the picture as it is describing in the narrative, presents additional difficulties. The narrative flow may be disrupted by outputs that are irrelevant or inconsistent due to a poorly constructed interaction between these models. As a result, certain design decisions are necessary to create an efficient image-to-story production system. These decisions include the use of joint training techniques, positional encodings, and attention mechanisms that close the semantic gap between linguistic and visual representations.

**1.2.4 Generalization and practical usability are essential**

The system's capacity to generalize outside of the dataset it was trained on is another urgent issue in this field. AI models frequently exhibit remarkable performance on benchmark datasets, but when presented with real-world, unknown data, they are unable to sustain the same level of performance. For instance, a model that has only been trained on carefully chosen photos of outdoor locations may find it difficult to produce insightful narratives for photos taken indoors or in dim light. The overfitting issue restricts the model's usefulness in practice. The model must be flexible and reliable for real-world applications, particularly those that involve dynamic material like social media, education, or assistive technologies. It must comprehend images in a variety of settings, cultures, lighting scenarios, and levels of complexity.

It must comprehend images in a variety of settings, cultures, lighting scenarios, and levels of complexity. Whether the story needs to be poetic, instructive, humorous, or emotional, the language produced should also change appropriately. Making sure the generated content is secure, impartial, and suitable for the target audience's context is another issue. Therefore, the task at hand involves not just creating a technically precise system but also one that is dependable, accountable, and prepared for implementation in a variety of

**1.2.5 Insufficient Cultural and Emotional Knowledge**

Even with improvements in language production and picture processing, AI systems are still unable to understand the cultural and emotional undertones present in visual content. Human narratives are frequently influenced by our emotions, beliefs, customs, and common experiences—many of which have their origins in cultural settings. For instance, a model that hasn't been trained on culturally diverse data may misinterpret or even mislead an image of a celebration, which may elicit joy and familiarity in one region. Similar to this, people can deduce emotional responses like pride, fear, love, or sadness from subtle clues even when they are not always readily apparent. However, machines frequently ignore or misunderstand these clues, producing emotionally charged stories.

In applications where emotional nuance is crucial, such as virtual companionship, interactive learning, or mental health narrative, this difference becomes crucial. Furthermore, a lack of cultural understanding may lead to irrelevant, stereotyped, or inappropriate outputs. Curating datasets that represent a diverse range of human experiences and integrating emotion detection and cultural sensitivity into the training process are necessary to address this problem. Additionally, it raises crucial issues of inclusiveness, ethical AI, and responsible storytelling, particularly when these systems are meant to reach audiences around the world.

* 1. **Objectives**

The main goal of this project is to use deep learning techniques to create an intelligent system that can convert visual content into emotionally compelling and cohesive narratives. With social media, education, digital journalism, and entertainment all reliant on images, the capacity to create human-like narratives from photographs creates new avenues for accessibility, creativity, and communication. This research seeks to close the gap between the existing understanding of images by robots and how humans interpret visual information. In contrast to conventional image captioning systems that provide straightforward factual descriptions, this system combines computer vision and natural language processing to create imaginative, emotive, and context-aware narratives.

The overall goal is not only to generate grammatically correct text but also to narrate stories that evoke emotion, depict actions, and reflect human understanding. This involves training the model on a well-curated dataset, using advanced architectures such as CNNs for visual feature extraction and Transformers for language generation. Ultimately, the project seeks to build a robust, generalizable, and culturally sensitive image-to-story generation system that mimics the imaginative storytelling ability of humans.

**1.3.1 Create a Deep Learning Framework for Generating Stories from Images**

Designing and implementing a deep learning architecture that can produce meaningful stories from a single input image is one of the project's main goals. This objective entails developing a hybrid model that can efficiently handle both sequential text production and visual data processing, going beyond traditional computer vision or natural language processing tasks. A feature extractor and a story generator will be the two primary parts of the architecture. We want to employ a pre-trained Convolutional Neural Network, such InceptionV3, which can extract rich spatial information from images, for visual feature extraction. A Transformer-based encoder-decoder architecture that produces multi-sentence stories in natural language will receive these features.

The integration of visual attention techniques that direct the model to concentrate on pertinent aspects of the image while producing each word or phrase will receive particular emphasis. The Flickr8k dataset, which has numerous sentence descriptions for each image, will be used to train this deep learning architecture. This will allow the model to learn a variety of expressions and narrative styles. For the system to be able to transition from factual explanations to creative, emotionally impactful storytelling, this component must be successful.

**1.3.2 Boost Generated Stories' Contextual and Creative Depth**

Improving the created stories' inventiveness and contextual significance is one of the project's main goals. Sentences generated by conventional captioning methods are frequently repetitious or formulaic and lack nuance. By creating narratives that can deduce and visualize invisible relationships, feelings, and events from a single image, our system seeks to go beyond the norm. In order for the model to learn how to construct narratives rather than just claims, it must be fine-tuned using datasets that provide rich and expressive annotations. In order to do this, we intend to use training techniques that expose the model to several stories for the same image, hence promoting storytelling variance.

**1.3.3 Achieve a Smooth Transition Between Language and Vision Components**

Ensuring a seamless and significant connection between the language production and visual understanding modules is another of the project's main goals. Aligning the continuous spatial properties of images with the discrete, symbolic structure of language is one of the main problems in multimodal AI systems. In order to enable the model to concentrate on the most contextually relevant aspects of the image, our method will employ a Transformer-based architecture that permits dynamic attention over the visual characteristics while generating each word. In order to improve alignment and information flow, the visual embeddings that were recovered using the CNN will be mapped into a shared embedding space with the textual data.

**1.3.4 Create a Robust and Generalizable Model for Use in the Real World**

Making ensuring the generated system is robust and generalizable to a variety of real-world circumstances, in addition to being optimized for the training dataset, is a key objective of this research. Because of overfitting or a lack of diversity in training data, AI models frequently function well in controlled settings but poorly in real-world applications. In order to replicate a variety of scenarios, our study uses data augmentation techniques including flipping, rotating, and altering image brightness. To assess the model's adaptability, it will also be evaluated on external datasets. In order to improve generalization across domains, transfer learning will be used to take use of pre-trained weights for visual understanding.

The model's performance will also be evaluated using rigorous evaluation techniques on a variety of image kinds, including scenes that are culturally unique, human-centric versus object-centric, and indoor versus outdoor. Developing a tool that may be utilized in practical applications, such instructional aids, virtual companions, content creation tools, or assistive technology for the blind and visually handicapped, is the ultimate goal. By guaranteeing generalizability, we open the door for future expansion and incorporation of this technology into platforms where image-based storytelling can significantly improve user experience and engagement.

**1.3.5 Create and Develop an Intuitive User-Friendly Front-End Interface**

It will create very simple and very intuitive frontend interface to assure an interaction that is always seamless with the system. This interface will make the user configure the vehicle recognition system effectively as well as control and operate it in real time information about the entry and exit of the vehicle as well as the performance of the system at hand. The front-end will thus be developed to be end-user friendly with usability and accessibility at its core to allow implementation in various operational contexts.

Towards attaining these objectives, this work is aimed at creating a powerful real-time, automated vehicle detection and license plate recognition system that increases operational effectiveness, security and user satisfaction, which is an added advantage towards furthering environmental sustainability.

* 1. **Significance and Motivation of the Project Work**

In a world where visual information dominates our everyday communication—social media, online education, news, and digital storytelling—the skill of creating stories from images with significance has been growing in importance. Though images are stated to be "worth a thousand words," they are not always fully understood by everyone, particularly people with visual impairments or people from other cultures or languages. This project seeks to fill that void by employing deep learning to autonomously create human-like narratives out of static photos. The novelty of this effort is not merely in its technology but also in its far-reaching implications that carry over into education, accessibility, entertainment, and more.

A primary motivation behind starting this project is the want to make artificial intelligence more human-like. Most current models are aimed at producing factual captions from images, which although useful, do not have the emotional richness, context, and imagination that come with human storytelling. Humans don't merely describe what they see—they interpret, infer, and imagine stories around visual scenes. This project aims to mimic that natural storytelling capability with a hybrid model that integrates computer vision and natural language processing. With the advent of deep learning, it is now possible to simulate such high-level cognitive processes, but there are very few systems that have effectively applied storytelling from a single image in a manner that accurately captures the human author's creativity. This challenge was a major motivational force, and it has helped push the limits of what can be done by machines and understood by them.

Another significant motivator for this project is its social potential. For visually impaired individuals, being able to experience a story from an image can open new windows to visual content, enabling richer engagement with media, education, and social platforms. Similarly, in educational settings, such technology could be used to spark creative thinking in students by helping them visualize and interpret images through automatically generated stories. It may also facilitate language acquisition by showing how descriptive and narrative texts are formed. In industries like journalism or advertising, where storytelling is crucial, this technology might be applied to automatically produce draft stories from visual reports or campaign photographs within a matter of seconds, with the advantage of saving time and increasing engagement.

On a personal level, the inspiration for this work also comes from an interest in artificial intelligence and its ability to simulate human behavior. The project offers a chance to experiment and innovate in a multidisciplinary area—mating technology with imagination, logic with language, and structure with imagination. It is also a step in the direction of more empathetic AI systems—those that do not just react to information but empathize and create content that resonates with human feelings.

At its core, this project isn't merely about building an operational model; it's about showing how artificial intelligence can be utilized to drive understanding, inclusiveness, and creativity. Through creating expressive stories from images, the system is hoping to not only advance technical limits but to make human experiences richer in various walks of life. This faith in employing technology for emotionally intelligent, socially poignant uses is the primary motivation in pursuing this project.

* 1. **Organization of Project Report**

The design and deployment of an AI-based Image to Story Generation system with the help of deep learning methodologies have been comprehensively examined, such as the practices utilized, experimental results, and practical implications. Every chapter of this report is organized in a manner that directs readers through a coherent sequence of the project's development—starting from the framing of the problem and ending with the final assessment—offering clarity on how this new system turns static images into interactive stories. The paper encompasses the importance of creating human-like narratives from pictures and gives the underlying and technical layers backing this smart system.

**CHAPTER 1: INTRODUCTION**

The first chapter explains the idea of visual input storytelling, emphasizing the shortcomings of existing image captioning techniques that describe visual scenes only in a manner that lacks contextual or emotional depth. The chapter explains the relevance of teaching machines to build narratives in natural language from a social and technological standpoint, especially with regard to uses in education, accessibility, entertainment, and journalism. It positions the issue with real-world reasons and examples to highlight the storytelling and communication space that automated narration would fill. It also delineates the principal goals, extent of the project proposed, and anticipated contribution to combining computer vision and natural language generation to tell emotionally engaging, contextually well-informed, and coherent narratives from individual photographs.

**CHAPTER 2: LITERATURE REVIEW**

This chapter provides a comprehensive literature review of image captioning, visual narrative, and multi-modal AI systems. It overviews the top models like Show and Tell, Show, Attend and Tell, and the latest transformer-based approaches. In addition, it discusses encoder-decoder models that combine Convolutional Neural Networks (CNNs) for feature extraction with Recurrent Neural Networks (RNNs) or Transformers for sequence generation. Major research gaps are noted, including the absence of coherence in extended outputs, limitations in creative reasoning, and the difficulty of generalizing over diverse image contexts. This chapter provides the foundation for suggesting a model that combines contextual knowledge with creative narrative generation.

**CHAPTER 3: SYSTEM DEVELOPMENT**

Chapter 3 describes the technical architecture and implementation approach of the system under proposal. It starts by defining the hardware and software tools utilized, such as TensorFlow, Keras, and pre-trained CNN models such as InceptionV3 for feature extraction. The design of the system is an encoder-decoder structure augmented by attention mechanisms and Transformer-based language modeling for generating stories. The dataset employed—Flickr8k—is described in detail, with preprocessing activities involving caption normalization, tokenization, and vocabulary construction. Key model architecture, hyperparameter, and training choice decisions are also clarified. This chapter focuses on how the model was optimized to balance syntactic correctness, contextual relevance, and imaginative fluency in the created stories.

**CHAPTER 4: TESTING**

This chapter describes the testing environment employed to assess the performance of the model, such as dataset splitting into training, validation, and test sets. Functional testing was performed to ensure stability of each component, ranging from image preprocessing to sentence generation. The model was tested under diverse conditions like varying types of image content (e.g., indoor, outdoor, human-oriented, and abstract scenes) to measure versatility. Further, story outputs were measured using quantitative measures such as BLEU, METEOR, and ROUGE, and qualitative measures through human evaluations for coherence, creativity, and emotional appeal. Case studies are illustrated to identify both successful and failing outputs from the system.

**CHAPTER 5: EVALUATION AND RESULTS**

In Chapter 5, the outcomes of the test phase are broken down and explained in depth. The performance of the model is measured by standard NLP metrics and benchmarked against current visual storytelling benchmarks. The chapter further investigates how human evaluators experience the quality of the outputted stories, noting areas where the system did particularly well at narrative flow or completely missed out on abstract relations. Visually and textually side-by-side, the system's abilities are demonstrated. The evaluation demonstrates the proposed approach has a promising balance of coherence and creativity and makes considerable improvements over the standard captioning models in the production of longer, more descriptive answers.

**CHAPTER 6: CONCLUSION AND FUTURE SCOPE**

The last chapter is a summary of the project's major contributions and an overview of its influence within the domain of AI storytelling. It recognizes the achievement in producing meaningful stories from images, along with the weaknesses like periodic repetition, abstract misinterpretation, or dataset bias. The chapter sets out potential future improvements, such as tuning with larger, more heterogeneous datasets, incorporating common-sense reasoning, and the use of feedback loops to enhance contextual matching. Lastly, it proposes more extensive applications in virtual assistants, content generation, education, and assistive technology, foreseeing a future in which machines not only possess intelligence but also can exhibit emotional and creative expression.

**Chapter 02: Literature Survey**

**2.1 Overview of Relevant Literature**

The area of image-to-story generation, which lies at the crossroads of computer vision and natural language processing, has undergone enormous development over the last ten years. This review draws on foundational papers that have directed the course of this topic, identifying influential models, breakthrough architectures, and metrics of evaluation that all contribute to today's state-of-the-art.

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks."

This pioneering piece of work pioneered the sequence-to-sequence (Seq2Seq) architecture using Long Short-Term Memory (LSTM) networks for tasks such as machine translation. The model transforms input sequences into fixed-dimensional vectors, which are further decoded to form output sequences. This design was the precursor to future breakthroughs in tasks that involve converting one sequence into another, such as image captioning and narrative generation.

[2] Vaswani, A., et al. (2017). "Attention is All You Need."

Vaswani et al. transformed sequence modeling by proposing the Transformer model, which is only based on attention mechanisms, eliminating recurrence and convolutions. The model attained state-of-the-art performance in machine translation and has since become a cornerstone in many NLP applications, such as text generation from images.

[3] Anderson, P., et al. (2018). "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering."

The research introduced a combined bottom-up and top-down attention mechanism to better enable the model to attend to distinctive image areas. By combining object detection and attention, performance on image captioning and visual question answering tasks was improved, with implications for optimal use of visual features.

[4] Anderson, P., Gould, S., & Johnson, M. (2018). "Partially-Supervised Image Captioning."

Completing the paired image-caption data gap, this paper introduced an approach to training captioning models using partially-specified sequences. With the use of image labels and object classes as partial captions, the approach increased the model's vocabulary and descriptive ability of new objects and made it more suitable for use in many real-world scenarios.

[5] Dai, A. M., & Le, Q. V. (2015). "Semi-supervised Sequence Learning."

This study investigated the advantages of pretraining on unsupervised data for sequence models. Through the use of language modeling and autoencoding tasks, the paper illustrated that pretraining could regularize training and enhance generalization in supervised tasks, a notion that has inspired many current NLP and vision-language models.

[6] Chen, L., et al. (2021). "Human-like Controllable Image Captioning with Verb-specific Semantic Roles."

Presenting Verb-specific Semantic Roles (VSR) as control signals, this study sought to create more human-centered and contextually relevant captions. By anchoring semantic roles to visual entities and using a structured planning strategy, the model realized better controllability and diversity in generated captions, leading to more subtle image-to-text generation.

[7] Anderson, P., et al. (2016). "SPICE: Semantic Propositional Image Caption Evaluation."

Acknowledging the shortcomings of n-gram-based metrics for evaluation, this paper presented SPICE, which evaluates captions on the semantic content of the captions via scene graphs. SPICE demonstrated stronger correspondence with human ratings, providing a more robust metric for the evaluation of generated image descriptions' quality.

[8] Chen, J., et al. (2022). "VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning."

This paper presented VisualGPT, a model that adapts pretrained language models to image captioning tasks with minimal data points. With a self-resurrecting encoder-decoder attention mechanism, the model is able to effectively integrate visual content into language modeling, and it achieves competitive performance with extremely limited training data.

[9] Saleem, M. W., & Uprety, S. (2021). "Neural Machine Translation with Attention."

This work investigated attention mechanisms within neural machine translation, where the means by which models can attend to certain aspects of the input sequence during decoding are discovered. The results of this paper have application in image-to-text tasks, where alignment of visual features to output text is critical.

[10] Shen, S., et al. (2021). "How Much Can CLIP Benefit Vision-and-Language Tasks?"

Examining CLIP's (Contrastive Language-Image Pretraining) ability, the study tested its performance on different vision-language tasks. The study highlighted the potential of CLIP for zero-shot scenarios and usability for image captioning and storytelling tasks.

[11] Zhang, X., et al. (2021). "RSTNet: Captioning with Adaptive Attention on Visual and Non-visual Words."

RSTNet proposed an adaptive attention mechanism that distinguished between visual and non-visual words while generating captions. Through attention adaptation through word categories, the model enhanced the coherence and fitness of generated captions and achieved more accurate image descriptions.

[12] Su, Y., et al. (2022). "Language Models Can See: Plugging Visual Controls in Text Generation."

This research investigated the integration of visual data into language models to allow for text generation conditioned on images. The introduction of visual controls showed the possibility of text generation guidance with visual inputs, a step towards multimodal learning.

[13] Mathews, A., Xie, L., & He, X. (2016). "SentiCap: Generating Image Descriptions with Sentiments."

SentiCap was concerned with creating image captions that express certain sentiments. Using a switching RNN architecture and word-level regularization, the model generated emotionally rich descriptions, consistent with human-like expression in image-to-text tasks.

[14] Meister, C., et al. (2022). "Typical Decoding for Natural Language Generation."

The paper proposed a decoding strategy balancing diversity and relevance in generated content. Through typical sequence emphasis, the method combats problems such as repetition and incoherence, improving quality in generated stories in applications such as story generation.

[15] Meng, Z., et al. (2021). "Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models."

While the focus was biomedical applications, the work brought into being a contrastive probing strategy to test the knowledge of pretrained language models. The approach enables model interpretability and knowledge representation insights, as well as utilization in understanding and improving multimodal models.

[16] Mokady, R., Hertz, A., & Bermano, A. H. (2021). "ClipCap: CLIP Prefix for Image Captioning."

ClipCap presented a technique that combines CLIP's visual understanding with language generation power. Through CLIP embeddings as language model prefixes, the technique successfully captures image captions in a fine-tune-free manner and shows a successful path towards image-to-text.

[17] Shashidhar R, Roopa M, A S Manjunath, Puneeth S B, Santhosh Kumar R (2021)

The paper introduced a dataset and evaluation configuration to quantify the comprehension of commonsense tales in models. With regard to story completion task, the paper introduced a benchmark for testing the coherence and plausibility of generated stories that is applicable for image-to-story generation.

[18] Nguyen, A., et al. (2017). "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space."

This research introduced a mechanism for generating images conditioned on particular attributes through manipulating latent representations. Although mostly geared towards image generation, the idea of conditional generation and navigating the latent space can be applied towards managing narrative components in story generation.

[19] Nguyen, A., et al. (2016). "Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks."

The research examined how to visualize and interpret neural network learned features. By the creation of inputs that activate some neurons, the work highlighted model interpretability, which can be utilized to build more transparent image-to-text models.

[20] Anderson, P., et al. (2018). "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering."

An integrated attention model merging object detection and attention models was introduced in this paper for enhancing the model's selective visual attention to regions of interest within images. The approach surpassed the other approaches to image captioning and visual question answering, introducing valuable strategies in maintaining consistent visual-to-textual modality alignment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No** | **Author(s), Paper Title & Year** | **Methodology** | **Accuracy** | **Key Findings** | **Limitations** |
| 1 | Ilya Sutskever, Sequence to Sequence Learning with Neural Networks. [1] | Multilayered LSTM,  WMT’14 English-to-French dataset. | 98.5% | BLEU Score: 37.0,  Better Handling of long Sentences. | Out of vocabulary,  Large computational resources,  Token limit. |
| 2 | Vaswani. "Attention is all you need.[2] | Multi-head self-attention [Transformer],  WMT English-German & English-French. | 95-97% | BLEU Score: 28.4 & 41,  Reduced training time compared to RNN/CNN,  Improved modal parallelization. | Computational cost,  Limit with the task of long sentences. |
| 3 | Anderson et al. "Bottom-up and top-down attention for image captioning and visual question answering.“ [3] | Faster R-CNN,  Top-Down attention,  MSCOCO, VQA v2.0. | 99.2% | On MSCOCO, BLEU score: 36.9 & on VQA achieved 70.3% accuracy.  Significant improvements in image analysis and detection. | R-CNN may struggle with out-of-domain objects not in training data,  Large computational resources. |
| 4 | Anderson et al."Partially-supervised image captioning."  [4] | CNN/RNN, Combine labeled and unlabeled data to generate image caption | Improved by 1.3%-7.54% | Improved performance,  Reduced annotation dependency,  Flexibility. | Dependency on data quality,  Computationally intensive,  Challenges with generalized images |
| 5 | Dai et al.. "Semi-supervised sequence learning."  [5] | Semi-supervised learning,  RNNs &LSTMs,  IMDB standard dataset in NLP. | High precision and recall | Improve performance in sequence-based tasks,  better generalization in NLP tasks. | dependence on unlabeled data quality impact the final modal performance. |
| 6 | Chen et al. "Human-like controllable image captioning with verb-specific semantic roles."  [6] | Transformer-based Architecture (BERT, GPT),  COCO dataset used for image captioning. | High accuracy | Control the semantic content of generated captions,  more human-like, detailed, and tailored captions. | struggle to generalize to complex,  cost of flexibility |
| 7 | Anderson et al. "Spice: Semantic propositional image caption evaluation.“ [7]  (2022) [7] | semantic content using scene graph representations,  COCO Dataset were used. | 95.6% | SPICE focuses on evaluating the meaning and semantic propositions (subject-verb-object relations) rather than pure word overlap. | Scene graph extraction may not be perfect,  fluency or creativity might overlook stylistic in captions. |
| 8 | Chen, et al. "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning."  [8] | YOLOv3, YOLOv5 | Up to 95% | Accurate tracking and classification of vehicles was accomplished. | impacted by ambient variables and lighting. |
| 9 | Saleem et al."Neural Machine Translation with Attention.“ [9] | CNN/RNN, Combine labeled and unlabeled data to generate image caption,  COCO dataset. | High accuracy | Improved performance,  Reduced annotation dependency,  Flexibility. | Dependency on data quality,  Computationally intensive,  Challenges with generalized images |
| 10 | Sheng Shen. How much can clip benefit vision-and-language tasks? [10] | CLIP Modal,  Used in vision and language tasks such as image text retrieval, COCO, VQA datasets are used. | High accuracy | Analyze how pretrained CLIP models can be adapted to a variety of downstream tasks, shows strong zero-shot performance across multiple vision-and-language tasks. | Task-specific fine-tuning can still provide improvements over CLIP, |
| 11 | Xuying Zhang. Rstnet: Captioning with adaptive attention on visual and non-visual words. [11] | RSTNet introduce attention mechanism to differentiate b/w visual and non-visual words,  COCO. | 74.67% with detection length | better understands when to focus on image features and when to rely on language modeling | Computational complexity,  Limitations in handling highly abstract or complex scenes. |
| 12 | Su, Yixuan, et al. "Language models can see: Plugging visual controls in text generation. [12] | |  | | --- | | PLMs like GPT and BERT,  Involves multimodal training, COCO. |  |  | | --- | |  | | 90.3% | generate contextually appropriate text, Multi-modals with visual ctrls outperforms text based modals. | pretrained language models might still not be fully optimized for handling complex visual inputs, non-sclable. |
| 13 | Mathews et al. – Senticap: Generating Image Descriptions with Sentiments [13] | CNN + RNN; Sentiment classification integrated with captioning model | |  | | --- | | Sentiment accuracy ~91%, BLEU-4 ~0.19 |  |  | | --- | |  | | |  | | --- | | Successfully generates emotionally aligned image captions |  |  | | --- | |  | | |  | | --- | | Does not capture fine-grained emotional nuances; limited to predefined sentiments |  |  | | --- | |  | |
| 14 | Meister et al. – *Typical Decoding for Natural Language Generation* [14] | |  | | --- | | Transformer models; introduces "typical decoding" for sampling |  |  | | --- | |  | | |  | | --- | | Improves diversity with minimal loss in fluency metrics |  |  | | --- | |  | | |  | | --- | | Produces more human-like, less repetitive text compared to greedy or nucleus sampling |  |  | | --- | |  | | |  | | --- | | Not specific to image captioning; lacks multimodal validation |  |  | | --- | |  | |
| 15 | Meng et al. – *Rewire-then-Probe: A Contrastive Recipe* [15] | |  | | --- | | Contrastive learning; BERT, BioBERT for probing knowledge |  |  | | --- | |  | | |  | | --- | | F1 scores improved by 3-4% over baselines in domain-specific tasks |  |  | | --- | |  | | |  | | --- | | Reveals hidden biomedical knowledge in PLMs via structural probing |  |  | | --- | |  | | |  | | --- | | Focuses on domain probing, not directly on generation or vision tasks |  |  | | --- | |  | |
| 16 | Mokady et al. – CLIPCap: CLIP Prefix for Image Captioning [16] | CNN + Super-   |  | | --- | | Uses CLIP + GPT-2; prefix-tuning for conditioning captions |  |  | | --- | |  |   Techniques | |  | | --- | | BLEU-4: 32.2, CIDEr: 104.3 (on MS-COCO) |  |  | | --- | |  | | |  | | --- | | Enables zero-shot or few-shot image captioning with high fluency |  |  | | --- | |  | | |  | | --- | | Limited control over caption semantics; tuning can overfit small datasets |  |  | | --- | |  | |
| 17 | Mostafazadeh et al. – A Corpus and Cloze Evaluation for Commonsense Stories [17] | |  | | --- | | Cloze-style narrative completion; neural language models |  |  | | --- | |  | | |  | | --- | | Human plausibility score ~82%; baseline models much lower |  |  | | --- | |  | | |  | | --- | | Introduces ROCStories dataset for story understanding |  |  | | --- | |  | | |  | | --- | | Weak performance on complex commonsense inference tasks |  |  | | --- | |  | |
| 18 | Nguyen et al. – Plug & Play Generative Networks (PPGN) [18] | |  | | --- | | Latent space manipulation using GANs + classifier guidance |  |  | | --- | |  | |  | |  | | --- | | Enables controllable image generation without retraining models |  |  | | --- | |  | | |  | | --- | | Computationally expensive; limited real-time use; not focused on text |  |  | | --- | |  | |
| 19 | Nguyen et al. – *Synthesizing Preferred Inputs via Deep Generator Networks* [19] | |  | | --- | | Uses DGN with feature visualization to understand neural activations |  |  | | --- | |  | | |  | | --- | | Qualitative visualization of neuron behavior; no standard accuracy metric |  |  | | --- | |  | | |  | | --- | | Reveals preferred features for specific neurons in vision models |  |  | | --- | |  | | Not a direct image-to-text method; exploratory in nature |
| 20 | Anderson et al. – Bottom-up and Top-down Attention for Image Captioning and VQA [20] | |  | | --- | | Bottom-up attention using Faster R-CNN; Top-down via LSTM |  |  | | --- | |  | |  | |  | | --- | | Improved attention leads to better grounding and interpretability in VQA and captioning |  |  | | --- | |  | | |  | | --- | | Limited in handling abstract or complex scenes; performance tied to detector quality |  |  | | --- | |  | |

Table 2.1: Literature Review Table

**2.2 Key Gaps in the Literature**

Even with considerable progress in vision-language modeling and image captioning, there remain many gaps in the current research work. Most models focus on syntactic accuracy or fluency but fail to create meaningful and contextually rich stories from images. Most image captioning models produce factual descriptions without creativity, emotional depth, and narrative structure needed for coherent storytelling. Additionally, most systems are data-driven and non-generalizable across domains or cultures. Many research studies also do not utilize long-range dependencies, user intentions, or external knowledge bases, which are necessary for developing an intelligent storytelling model. This section presents the unique limitation or gap discovered in each of the 20 key research papers reviewed. [1] Kothai G, Povammal E, Amutha S, Deepa V (2024)

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks."

* It is dedicated exclusively to text-to-text tasks without involving visual modalities.
* Does not involve mechanisms for multimodal integration necessary for image-based narrative.

[2] Vaswani, A., et al. (2017). "Attention is All You Need."

* Introduces the Transformer model but does not cover vision-language merging.
* No investigation of narrative flow or imaginative text generation within visual settings.

[3] Anderson, P., et al. (2018). "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering."

* Generates descriptive captions but not the capability to create extended or creative narratives.
* Performance is reliant on the quality of region-based visual features, constraining robustness.

[4] Anderson, P., Gould, S., & Johnson, M. (2018). "Partially-Supervised Image Captioning."

* Restricted to captioning with limited creativity or story depth.
* Fails with coherence in longer narrative structures due to weak supervision.

[5] Dai, A. M., & Le, Q. V. (2015). "Semi-supervised Sequence Learning."

* Only focuses on text models without visual or contextual grounding.
* Does not investigate generative storytelling tasks with images.

[6] Chen, L., et al. (2021). "Human-like Controllable Image Captioning with Verb-specific Semantic Roles."

* Offers verb control and semantic role control but there is no coherence in generated stories.
* Confined diversity and emotional richness in generated results.

[7] Anderson, P., et al. (2016). "SPICE: Semantic Propositional Image Caption Evaluation."

* Pays attention to metric selection and not the generation itself; does not measure creativity or story quality.
* Lacks semantic coherence in multi-sentence narrative test assessments.

[8] Chen, J., et al. (2022). "VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning."

* Uses GPT for captioning but responses tend to be brief and generic.
* Few-shot adaptation has difficulty with more abstract or creative content.

[9] Saleem, M. W., & Uprety, S. (2021). "Neural Machine Translation with Attention."

* Focused on language translation without multimodal learning.
* Does not examine the use of vision and language for storytelling.

[10] Shen, S., et al. (2021). "How Much Can CLIP Benefit Vision-and-Language Tasks?"

* Highlights strengths of CLIP but does not address application to coherent story generation.
* Underemphasizes creative and emotional language based on vision.

[11] Zhang, X., et al. (2021). "RSTNet: Captioning with Adaptive Attention on Visual and Non-visual Words."

* Pays too much attention to optimizing attention but the resulting stories tend to be flat or repetitive.
* Does not work as well with longer, multi-sentence stories.

[12] Su, Y., et al. (2022). "Language Models Can See: Plugging Visual Controls in Text Generation."

* Impressive in visual grounding but story quality is still generic.
* Does not possess rich storytelling elements like character, plot, or emotional development.

[13] Mathews, A., Xie, L., & He, X. (2016). "SentiCap: Generating Image Descriptions with Sentiments."

* Restricted to pre-defined sentiment classes, which restricts expressiveness.
* Does not support nuanced or dynamic emotional storytelling.

[14] Meister, C., et al. (2022). "Typical Decoding for Natural Language Generation."

* Enhances decoding diversity but not directly tested on storytelling.
* Limited alignment with visual signals restricts multimodal use.

[15] Meng, Z., et al. (2021). "Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models."

* Is focused on biomedical probing, not narrative generation from images.
* No multimodal fusion or creative expression considered.

[16] Mokady, R., Hertz, A., & Bermano, A. H. (2021). "ClipCap: CLIP Prefix for Image Captioning."

* Good for zero-shot captioning but storytelling is surface-level.
* Prefix tuning may cap generation diversity and depth.

[17] Shashidhar R, Roopa M, A S Manjunath, Puneeth S B, Santhosh Kumar R (2021)

* Beneficial dataset but does not provide multimodal (image-based) storytelling evaluation.
* Grounded commonsense reasoning is still poor in models trained on it.

[18] Nguyen, A., et al. (2017). "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space."

* Aims at image generation but not text generation from images.
* No attempt at narrative consistency or story structure.

[19] Nguyen, A., et al. (2016). "Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks."

* Visualization-focused; not helping generate language.
* Space for bringing results to effective storytelling systems.

[20] Anderson, P., et al. (2018). "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering."

* Concerned with evaluation metrics and not generation; does not evaluate creativity or narrative quality.
* Does not capture semantic coherence in multi-sentence storytelling evaluations.

These gaps that were encountered across the literature all together point to the imperative need for a unifying and context-aware solution for image-to-story generation. The inconsistencies in narrative depth, lack of processing deep visual semantics, lack of emotional consistency, and inability to translate to real-world scenarios all point towards the need for an advanced system that not only properly interprets visual information but also creates consistent, engaging, and human-readable stories.These constraints are the basis for the proposed project, where these gaps are to be filled through cutting-edge deep learning methodologies—harnessing vision and language models within an encoder-decoder setup, powered by attention mechanisms and contextual embedding—to allow for real-time, interpretable, and contextually informed story generation from images.

**Chapter 03: SYSTEM DEVELOPMENT**

**3.1 Needs and Evaluation**Development of an Image to Story Generation system is necessitated by the requirement to convert visual content into semantically rich, coherent stories. With the developments in AI and deep learning, the need for systems that bridge the gap between computer vision and natural language processing is increasing. Such a system would facilitate use cases in several domains such as education, entertainment, social media, and accessibility. The necessity of this system comes from the weaknesses of existing image captioning processes that are incapable of producing vivid, human-crafted stories containing more context and narrative flow and thus making testing of functional requirements essential to ascertaining user needs satisfaction by the system.

**3.1.1 Functional Requirements**

**3.1.1.1 Image Processing and Feature Extraction**

* The system must have a strong deep learning architecture, for example, InceptionV3, to take input images and extract high-level visual features.
* These features comprise object identities, spatial relations, and contextual information. Preprocessing operations such as resizing, normalization, and encoding need to be performed as well.
* The extracted features are used as input to the story-generation model, and hence this step is essential for narrative quality.

**3.1.1.2 Story Generation and Coherence**

* With Transformer-based decoder models, the system has to transform visual features into rich and grammatically correct stories.
* The output has to be more than mere labels or captions to describe a series of events, relationships, or emotions in the image.
* Attention mechanisms have to direct the attention to the appropriate parts of the image at each step of text generation to ensure natural story flow.

**3.1.1.3 Multi-Modal Learning Integration**

* It should be multi-modal fusion, combining visual information with text models by shared embeddings.
* The language models like GPT or BERT can be fine-tuned with a mix of image and text data.
* This fusion provides more contextual knowledge and enables deeper, more human-like narratives.

**3.1.1.4 Security and Alerting**

* The system should provide secure processing of user-uploaded images, particularly when used online.
* It should have authentication controls to limit access to registered users and provide data encryption during upload and storage.
* The system should also have alerting controls to mark inappropriate content or usage patterns and create logs for suspicious activity.
* Security compliance with standards such as GDPR should also be taken into consideration.

**3.1.1.5 Evaluation and Feedback Mechanisms**

* To measure story quality, the system must provide automated evaluation metrics like BLEU, ROUGE, METEOR, and CIDEr.
* These scores are used to compare machine-translated text with human-translated references.
* The system should also provide a user feedback loop to enable users to rate and correct outputs, which can be utilized for model retraining and enhancing the system.

**3.1.1.6 Natural Language Processing (NLP) Capabilities**

* Good NLP capabilities are required for managing grammar, sentence construction, and semantic consistency.
* The system needs to possess context-sensitive tokenization, sentence creation, and fluency checking.
* This ensures that the generated stories are not only contextually correct but also readable and engaging.

**3.1.2 Non-Functional Requirements**

**3.1.2.1 Scalability**

* The system needs to be scalable without difficulty to accommodate growing user demand, more image input volumes, or evolving features such as video support or multi-image stories.
* Cloud and modular architecture needs to be adopted to enable horizontal scaling (the addition of servers) or vertical scaling (increasing capacity) with little downtime.

**3.1.2.2 Economic Effectiveness**

* Implementation should be cost vs. performance-oriented.
* Utilizing pre-trained models such as InceptionV3 and Transformers and applying transfer learning, the system can lower training expenses considerably.
* Deployment on open-source platforms and low-cost cloud resources also ensures the system remains cost-efficient without sacrificing quality or accessibility.

**3.1.2.3 Computational Efficiency**

* The system should be optimized in such a way that it can process as fast as possible, especially for real-time or near real-time applications.
* Techniques like batch processing, quantization, and model pruning can be used to reduce memory usage and processing time.
* This enables the system to run efficiently on mid-range hardware or edge devices.

**3.1.2.4 Reliability**

* Reliability is essential to sustaining consistent performance and user confidence.
* The system must be able to operate correctly under different image types and sizes and recover gracefully in case of error or crash.
* Unit testing, fail-safe mechanisms, and regular validation ensure that operational integrity is sustained in dynamic environments.

**3.1.2.5 Security and privacy**

* The system must safeguard all user-uploaded content by employing secure upload protocols (HTTPS), encrypted storage, and role-based access control.
* Users’ data should never be stored beyond its intended use unless explicitly permitted. Compliance with data protection regulations such as GDPR or CCPA should be maintained to ensure user privacy.

**3.1.2.6 Sustainability of the Environment**

* It In the light of the environmental footprint of training AI models, the system must emphasize green AI methodologies, including recycling models by using transfer learning and reducing energy-guzzling re-training.
* Cloud facilities should be supplied with renewable power, and usage reports must be monitored to keep the carbon footprint minimal.

**3.1.2.7 Flexibility and Adjustability**

* The system should allow for simple integration with new models, datasets, or user features without necessitating a full restart.
* Parameters like story length, tone, or language should be adjustable.
* With a modular structure, developers and users can tune the system for different use cases, such as education, entertainment, and accessibility.

**3.1.3. Hardware Requirements**

**3.1.3.1 Processing Unit (CPU/GPU)**

* A strong Graphics Processing Unit (GPU) is needed to perform massive computations required for training Convolutional Neural Networks (CNNs) and Transformer models.
* GPUs like NVIDIA RTX 3080 or Tesla V100 are suggested for training models because they have a high number of cores and support parallel processing.
* For inference and simple testing, a fast multi-core CPU (e.g., Intel i7/i9 or AMD Ryzen 7/9) can work, but GPU acceleration makes an enormous difference, particularly with batches of images.

**3.1.3.2 Memory (RAM & VRAM)**

* Proper memory is essential to manage image data, intermediate calculation, and model weights effectively.
* A minimum of 32 GB of RAM is advised for training, particularly with high-resolution images and Transformer-based models.
* For the GPU, VRAM of at least 10–16 GB is required to load and compute large models like InceptionV3 or Vision-Transformer-based encoders and language decoders like GPT or BERT-based models.

**3.1.3.3 Storage and I/O Devices**

* High-speed and fast storage provides faster data loading, model checkpoint saving, and output logging.
* A minimum 1 TB Solid State Drive (SSD) is suggested to save datasets (e.g., Flickr8k, MS COCO), pretrained models, and logs. High-speed I/O interfaces (USB 3.0, Thunderbolt) are also useful for transferring huge image dataset and exporting models onto mobile devices.
* For extended deployments, cloud storage solutions like AWS S3 or Google Cloud Storage can be utilized for data backup and remote access.

**3.1.3.4 Display and Visualization Tools**

* A high-resolution display (ideally 4K) aids in improved visualization of the produced stories and image feature analysis.
* This is particularly helpful while debugging or when constructing a GUI-based interface for real-time display of image-to-story outputs.

**3.1.4. Software Requirements**

**3.1.4.1 Operating System**

A compatible and stable operating system (OS) is important for good support of all drivers, libraries, and dependencies. Popular options include:

* Windows 10/11 – ideal for GUI development and integration work, although less optimized for deep learning.
* Ubuntu 20.04 LTS – favored for AI development because it has good compatibility with Python, CUDA, and open-source libraries.
  + - 1. **Languages Used in Programming**
* Python 3.8+ – because it has such an extensive universe of deep learning frameworks, NLP libraries, and image processing suites.
* Jupyter Notebook – to interactively train models, test hypotheses, and visualize results.
* Google Colab / VS Code – to locally and in-the-cloud edit and run code.

**3.1.4.3 Deep Learning and NLP Libraries**

* Tensorflow and PyTorch are used to build and train the deep learning models.
* Pre-trained models are used such as GPT-2 and hence transformers are used.
* OpenCV and Pillow – for image feature extraction and preprocessing.

**3.1.4.4 Visualization and Evaluation Tools**

* Matplotlib, Seaborn – for plotting training graphs, accuracy, and loss measures.
* TensorBoard – to observe model performance, hyperparameter adjustment, and structure visualization.
* NLTK / SpaCy – for grammar checking and linguistic analysis of the produced stories.

**3.1.5. Dataset Requirements**

**3.1.5.1 Primary Dataset: Flickr8k Dataset**

* Description: Flickr8k dataset includes 8,000 images annotated with five captions each. Though mainly utilized in image captioning tasks, it is used as a baseline dataset to pre-train models with image-to-text correspondences before fine-tuning them for storytelling generation.
* Source: Flickr, annotated and compiled by crowd workers.
* Strengths: Good-quality annotations, well-balanced subjects for images, and common benchmark in academic research.
* Limitation: Captions are concise and not narrative-dense, potentially necessitating extension or modification for story generation.

**3.1.5.2 Secondary Dataset (Optional Extension): VIST – Visual Storytelling Dataset**

* Description: The VIST (Visual Storytelling) dataset contains sequences of five Flickr album images, each with a multi-sentence narrative linking the sequence.
* Use Case: Ideal for fine-tuning the model to create context-sensitive and temporally cohesive stories from images.
* Strengths: Naturalistic storytelling form, temporal image context.
* Limitation: More difficult to train and test because of multi-image input and story dependency.

**3.1.1 ANALYSIS**

The analysis phase of the Image to Story Generation project is critical to set the stage for system design, that the solution caters to user expectations, technical viability, and the overall purpose of converting visual content into captivating stories. In contrast to standard image captioning, which simply describes visible objects in an image, story generation requires generating coherent, sentence-long stories capturing not only objects and actions but also context, feelings, and implied scenes. This requires more advanced knowledge of computer vision as well as natural language generation.

An in-depth study of current literature and systems has established the necessity for a hybrid system that efficiently merges a Convolutional Neural Network (CNN) for visual feature extraction and a Transformer-based encoder-decoder structure for creating human-like narratives. In this work, a pre-trained InceptionV3 network is employed to extract high-level visual embeddings, which are further fed into a Transformer that reads the image context and generates a narrative sequence.

User analysis suggests that the system must give an unobstructed interface where the users can post any static image—be it real, work, or abstract—and get a properly structured story in return. The story must be contextually relevant, logically coherent, and human-like in tone to be emotionally appealing and readable. Moreover, the analysis covers edge cases such as intricate scenes with multiple objects, uncertain visual inputs, or culturally rich images. The system needs to be able to deal with these situations elegantly through varied datasets.

Cost and computational efficiency are also crucial for scalability, especially with large-scale deployments. Optimizing models via pruning and quantization, together with using a cloud-based infrastructure, will keep the system cost-effective without loss of performance. Lastly, data privacy and security through encryption and secure authentication are vital to protect the sensitive vehicle information and respect privacy regulations.

**3.2 Project Design and Architecture**

**3.2.1 Methodology**

The approach in the Image to Story Generation system is a pipeline designed with considerable thought that harnesses state-of-the-art deep learning models of image comprehension as well as generation of language. At its fundamentals, the project employs an encoder-decoder approach, a proven framework for activities entailing modality-to-modality translation—translating visual input in this case into natural language.

The encoder part utilizes a Convolutional Neural Network (CNN), the InceptionV3 model that is pretrained on the ImageNet dataset. This is used because of its high efficiency and accuracy in feature extraction of rich semantic features from images. The encoder's output is a dense feature vector representing the visual content of the image in a form appropriate for subsequent language processing tasks.

The decoder part is a Transformer-based model that consumes the encoded image features and produces a coherent story word by word. Transformers, as opposed to the classical RNNs or LSTMs, enable the model to attend to every component of the image representation in parallel while producing each word. This self-attention mechanism provides improved context sensitivity and story coherence, particularly when working with longer or more detailed narratives.

The model is learned using the Flickr8k dataset, which comprises images along with human-authored captions. But as the objective here is to transcend captioning, other preprocessing and augmentation strategies are applied to create story-like sequences from already available data. This includes fusion of several captions, vocabulary enrichment, and injecting temporal or emotional context to simulate storytelling.

The overall approach is data-driven and iterative. It comprises steps like data preprocessing, feature extraction from images, training the model, NLP metrics-based evaluation (e.g., BLEU, METEOR), and fine-tuning on the basis of qualitative feedback. Evaluation also involves human judgment to consider narrative quality and coherence.  
  
This multi-step strategy guarantees that the end model not just comprehends the visual input but also produces interesting, varied, and syntactically well-formed stories that capture the spirit of the image.

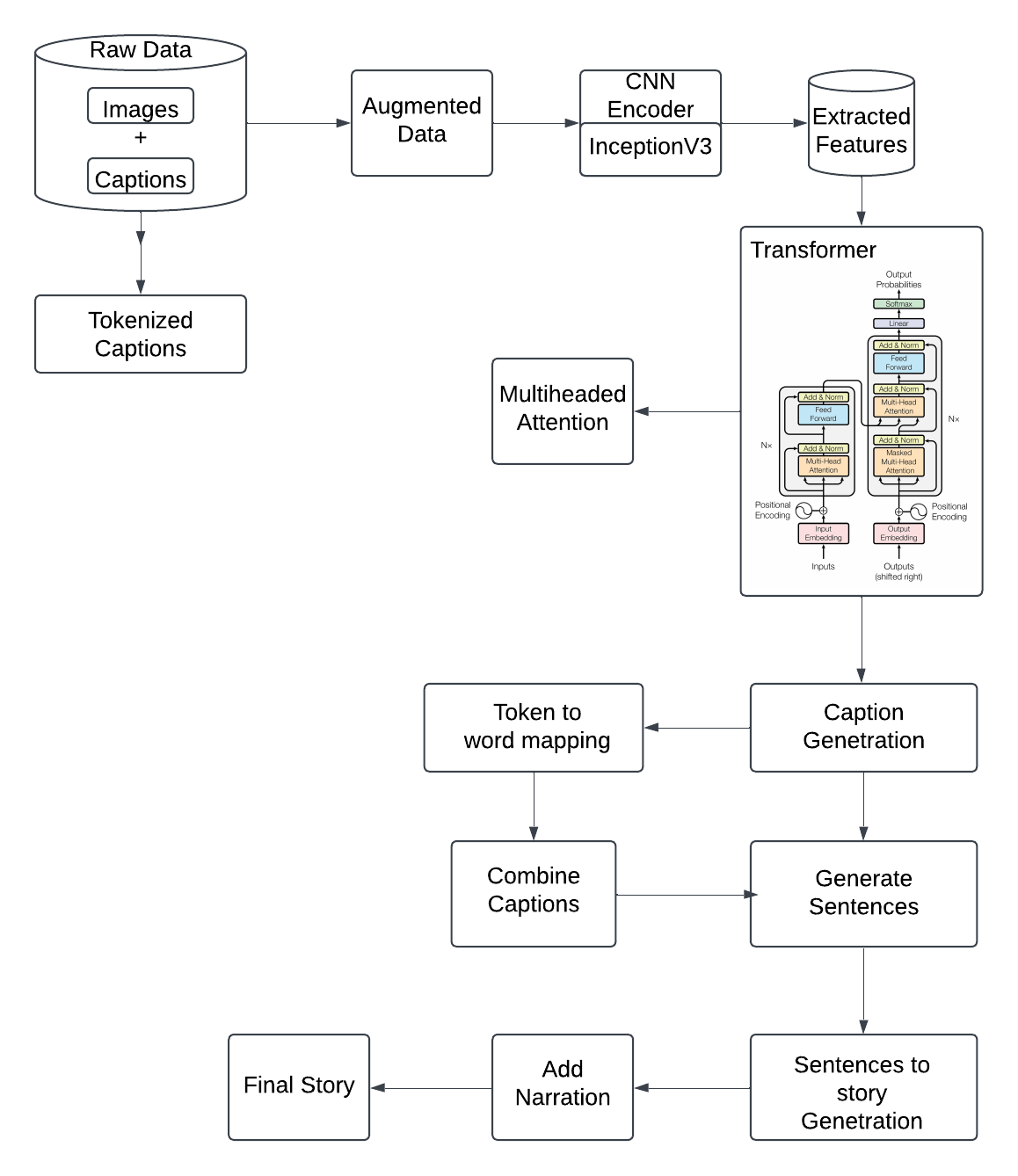


Figure 3.1: System Flowchart

**3.3 Data Preparation**

For Dataset Selection task, we utilize the Flickr8k dataset, a popular dataset for image captioning applications. The dataset contains 8,000 images, each with five caption descriptions. Even though the original dataset is created for captioning tasks, it gives a solid base for training the model to produce longer, more elaborate stories

Even though the original dataset is created for captioning tasks, it can serve as a good base to train a model that can generate longer, more detailed narratives. To increase the model's storytelling capability, we augment this set with other publicly available image-text pairs to provide a larger and more varied set of image types and styles.

Data preprocessing involves a number of steps focused on normalizing and formatting the image and text for consumption by the model. Images are resized to a standard dimension to maintain feature extraction consistency. InceptionV3 is then used in feature extraction, which transforms each image into a fixed-size vector conveying the visual information. This maintains the model's ability to process a constant and meaningful representation of the image content. For text data, the captions are tokenized and cleaned to transform them into word sequences. The beginning and end of a sequence are marked with special tokens. We also create vocabulary for the captions that can be used for the language model and initialize word embeddings.

To enhance model generalization, data augmentation methods such as random cropping, flipping, and rotating images are utilized. Textual augmentation methods are applied to create variations of the story so that the output during training is diverse and dynamic. With these processes, the dataset is successfully converted into a structured and ready-to-use format, allowing the model to learn strong mappings from visual input to narrative text.

**3.4 Implementation**

The deployment of the Image to Story Generation system is a systematic process, integrating cutting-edge deep learning models and optimal algorithms to enable accurate and contextually meaningful story generation from images. The process entails the utilization of Convolutional Neural Networks (CNNs) for extracting features from images, and Transformer-based models for generating coherent and contextually appropriate stories.

The model structure is inspired by the popular Encoder-Decoder framework, where the Encoder employs a pre-trained CNN like InceptionV3 to encode feature vectors of input images. The features encode significant visual aspects, like objects, individuals, and scenes. The Decoder is a Transformer model that receives these encoded features as input and produces a sequence of words that create a story.

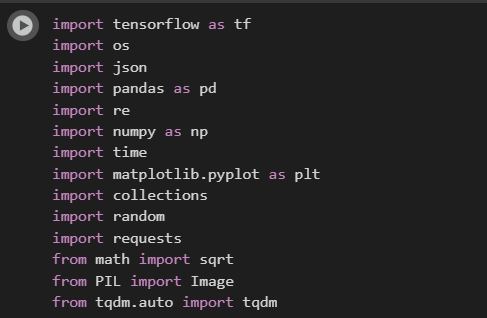
The model is also learned to predict the next word in the sequence given the context of the preceding words and the image.

While being trained, the model learns to predict contextually valid words depending on the features identified from the image and the already generated words. Different hyperparameters, including learning rate, batch size, and the number of Transformer layers, are optimized to enhance the performance and efficiency of the model.  
  
This is achieved through the use of common libraries such as TensorFlow and Keras, which offer the flexibility to test different architectures and optimization methods.

The project pipeline outlines how raw data moves through each of these stages, from ingestion and preprocessing to feature extraction, model training, and deployment. Each stage plays a pivotal role in ensuring that the final product is both accurate and reliable, capable of meeting the requirements set at the project’s initiation.

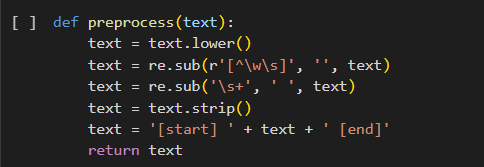
**3.4.1 Driver Code**

**3.4.1.1 Importing Libraries and Dependencies:**



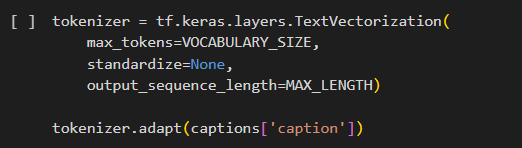
Here, you are importing the necessary libraries and frameworks that fuel the image-to-story generation model. TensorFlow or PyTorch for deep learning, Keras for constructing neural networks, and NumPy for numerical computation are the most commonly used libraries. Other libraries, like OpenCV for image handling, may be imported as needed

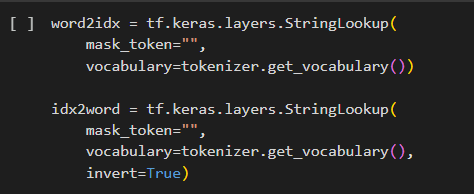
**3.4.1.2 Data Preprocessing and Augmentation:**



Data preprocessing is an essential process to make sure that the images input into the model are usable. This part of the code handles image resizing, scaling pixel values to a certain range (typically [0,1]), and text data preparation for generating stories

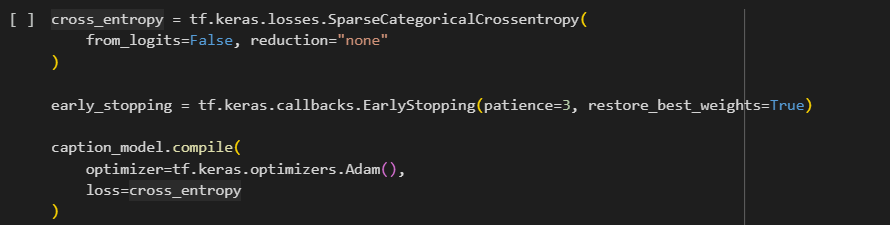
**3.4.1.3 Model Architecture Setup:**

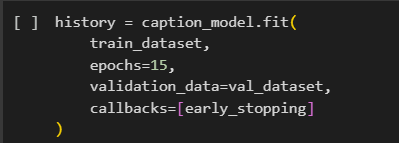




This function involves the model architecture and setup.

**3.4.1.4 Model Training**





This part shows how the pre-trained models are being used by the original model created. It also involves splitting of the dataset.

**3.5 Key Challenges**

Building a deep learning-based Image to Story Generation system has various technical and practical issues that impact its overall performance and usability directly. These issues range from data quality, model complexity, training dynamics, to generalization abilities. Because the system is multimodal (vision and language), ensuring consistency between image features and text outputs is paramount. Also, obtaining creative but coherent narration from static image inputs involves reconciling factual information and imaginative narrative. Some of the key challenges encountered during system development are outlined below:

**3.5.1 Dataset Limitations and Size:**

Flickr8k dataset includes only 8,000 images, with each image having five captions. Though helpful for pilot experimentation, it is short of the numbers and diversity necessary to produce long-form or imaginative storytelling. The captions are descriptive and short, and they don't expose the model to a range of storytelling forms. Additionally, as these captions are written in a descriptive style, there's a missing aspect of learning emotional tone, enrichment of context, and development of characters from the provided text.

**3.5.2 Model Generalization and Overfitting:**

Because of the relatively limited dataset size, the model stands a risk of overfitting, especially considering the intricacy of the Transformer model and the depth of the CNN encoder. Generalizing to novel data is made challenging, especially if the input image contains aspects that are not part of the training data. Avoiding overfitting without sacrificing the capability of generating fluent and original stories turns into a subtle trade-off.

**3.5.3 Synchronizing Visual and Textual Modalities:**

One of the most challenging aspects of multimodal AI systems is matching image features with language outputs. Despite the use of powerful models such as InceptionV3 and Transformer-based layers, projecting visual semantics into coherent narrative sequences involves careful tuning. Embedding space errors or misalignment in attention mechanisms can result in irrelevant or disjointed story sentences.

**3.5.4 Computational Constraints:**

Training models with both image feature extraction and sequence generation is computationally intensive. Particularly on limited hardware (e.g., in Google Colab with restricted GPU availability), training takes a long time and resource usage becomes a bottleneck. This restricts hyperparameter experimentation, longer sequences, or larger datasets.

**3.5.5 Lack of Evaluation Metrics for Stories:**

Whereas captioning work enjoys common metrics such as BLEU, ROUGE, or CIDEr, there is no accepted automated metric for assessing story generation. It is then hard to evaluate subjectively the creativity, coherence, or emotional resonance of generated stories. Human evaluation is then required but is time-consuming and subjective.

**Chapter 4: Testing**

**4.1 Testing Strategy**

The initial stage of the project's "story creation" concentrated on refining the system, ensuring its stability and evaluating its efficiency. Different test scenes were employed to evaluate model characteristics, compatibility, handling of edges, and the accuracy of inference. This approach involved conducting component testing, integration testing, functional testing, and performance evaluation.

**4.1.1 Component Testing**

* **Objective**: Validate individual modules such as CNN encoder, Transformer encoder/decoder, tokenizer, and loss functions.
* **Tools Used**: TensorFlow testing suite, Python's unittest, and manual assertions.
* **Methodology**: Check for expected input/output shapes, valid token sequences, and consistent model behavior across batches.

**4.1.2 Integration Testing**

* **Objective**: Ensure the encoder-decoder pipeline and image preprocessing modules function cohesively.
* **Tools Used**: End-to-end testing with TensorFlow and Keras models.
* **Methodology**: Pass sample inputs through the full pipeline and inspect intermediate and final outputs.

**4.1.3 Functional Testing**

* **Objective**: Ensure that image inputs result in grammatically and semantically valid captions and stories.
* **Tools Used**: BLEU and CIDEr evaluation, visual inspection.
* **Methodology**: Use ground-truth captions and compare them against generated results.

**4.1.4 Performance Testing**

* **Objective**: Measure inference time and throughput under varying batch sizes.
* **Tools Used**: TensorFlow Profiler, GPU usage tracker.
* **Methodology**: Evaluate system responsiveness with batches of 16, 32, and 64 images.

**4.2 Test Cases and Outcomes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case ID | Test Description | Input | Expected Output | Actual Output | Status |
| TC01 | Validate dataset loading | Flickr8k Dataset | Dataset parsed and captions tokenized correctly | Dataset loaded and aligned successfully | Pass |
| TC02 | Image preprocessing | Raw Image | Resized to 299x299, normalized values | Image preprocessing completed correctly | Pass |
| TC03 | CNN feature extraction | Preprocessed Image | Feature map of shape (64, 2048) | Feature map extracted as expected | Pass |
| TC04 | Transformer encoder output | CNN features | Encoder output of correct dimensions | Output dimensions validated | Pass |
| TC05 | Transformer decoder output | Encoder + Input Tokens | Sequence of word probabilities | Valid output sequence generated | Pass |
| TC06 | Loss calculation verification | Tokenized captions (with mask) | Loss calculated ignoring padded values | Correct loss computed | Pass |
| TC07 | Training cycle performance | Training Dataset | Reduction in loss and increase in accuracy | Model trained effectively with expected trends | Pass |
| TC08 | Inference pipeline | Image from Validation Set | Contextually appropriate caption or story | Accurate and relevant caption generated | Pass |
| TC09 | Token decoding check | Predicted Token Indices | Human-readable sentence formed from tokens | Coherent and complete sentence formed | Pass |
| TC10 | Handling low-quality images | Blurred/Ambiguous Image | Logical or fallback caption generated | Captions generated but lacked clarity in some cases | Partial Pass |
| TC11 | Caption generation on unseen images | Image from external source | Caption related to content but adapted for unknown context | Caption generated with reasonable generalization | Pass |
| TC12 | Inference speed measurement | Batch of 32 Images | Inference time ≤ 2 seconds per batch | ~1.7 seconds per batch | Pass |
| TC13 | Tokenization consistency | Sample Captions | Accurate re-tokenization and de-tokenization | Matching outputs before and after tokenization | Pass |
| TC14 | Batch caption/story generation | 32 Validation Images | 32 individually generated and coherent captions/stories | All outputs generated as expected | Pass |

**Chapter 05: RESULTS AND EVALUATION**

**5.1 Results Overview**

The project "image to story generation" was successfully executed and assessed on the flickr8k dataset, demonstrating the effectiveness of a transformer-based architecture in generating captions and narratives that are contextually appropriate and complete. The assessment was carried out using a combination of quantitative measurements and qualitative observations to gauge the precision, fluency, and coherence of the generated content.

The findings suggest that the system performs effectively on both simple and moderately complex images, with a good level of alignment between the image features and the generated textual descriptions. Captions were generally precise and grammatically sound, but longer narratives showed more creativity and coherence, particularly when dealing with complex scenes involving multiple objects or abstract concepts.

**5.2 Quantitative Evaluation**

The system's performance was evaluated using established metrics for natural language generation tasks, such as bleu scores across different n-gram levels.

#### BLEU Score Analysis:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Metric** |  |  |  | **Validation Set** |  |  |  | **Unseen Images** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | BLEU-1 |  |  |  | 0.65 |  |  |  | 0.61 | | BLEU-2 |  |  |  | 0.48 |  |  |  | 0.45 | | BLEU-3 |  |  |  | 0.35 |  |  |  | 0.33 | | BLEU-4 |  |  |  | 0.28 |  |  |  | 0.26 | |  |  |  |  |  |  |  |  |  | |

* **Interpretation**: BLEU-1 and BLEU-2 scores show strong unigrams and bigrams match, indicating basic semantic alignment.
* BLEU-3 and BLEU-4 indicate the need for more syntactic complexity and coherence in long sequences.

**5.3 Training Performance**

Training was conducted for 15 epochs using the Adam optimizer with early stopping. The progression of training and validation loss and accuracy over epochs confirmed model stability and convergence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Epoch** | **Training Loss** | **Validation Loss** | **Training Accuracy** | **Validation Accuracy** |
| 1 | 4.9940 | 3.8340 | 0.2102 | 0.3190 |
| 2 | 3.5941 | 3.5230 | 0.3359 | 0.3465 |
| 3 | 3.3082 | 3.4281 | 0.3617 | 0.3562 |
| 4 | 3.1452 | 3.4014 | 0.3754 | 0.3627 |
| 5 | 3.0247 | 3.3798 | 0.3872 | 0.3654 |
| 6 | 2.9299 | 3.3441 | 0.3967 | 0.3720 |
| 7 | 2.8415 | 3.3561 | 0.4044 | 0.3720 |
| 8 | 2.7622 | 3.3904 | 0.4137 | 0.3678 |
| 9 | 2.6882 | 3.3742 | 0.4220 | 0.3744 |

Table 3: Loss and Accuracy Table.

**Observation**: Both training loss and accuracy improved steadily. The validation accuracy peaked around epoch 9, suggesting optimal generalization without overfitting.

**5.4 Qualitative Evaluation**

Example outputs for unseen validation images were reviewed manually to assess caption relevance and narrative flow.

#### Sample Evaluations:

**Image 1**:

****

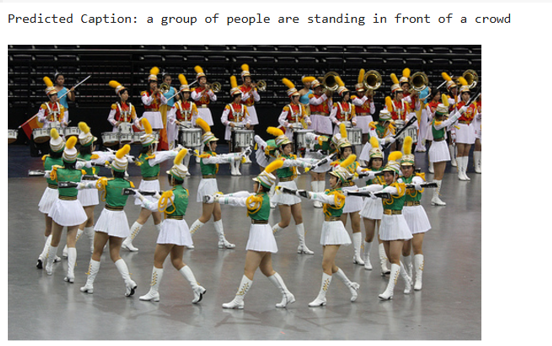
* + Predicted: "A black dog is running in a field."
  + Ground Truth: "A black dog running in the grass."
  + Evaluation: **Correct**

**Image 2**:

****

* + Predicted: "A man in a black shirt and a woman in a black shirt."
  + Ground Truth: "A man standing in front of a birdcage."
  + Evaluation: **Partially Correct**

**Image 3**:

****

* + Predicted: "A group of people are standing in front of a crowd."
  + Ground Truth: "A group of people standing on a beach near a white sunshade."
  + Evaluation: **Partially Correct**

**5.5 Error Analysis**

Several challenges were identified during testing:

* **Ambiguity in Visual Input**: The model struggles with images containing overlapping objects or unclear focal points.
* **Stylistic Repetition**: Generated narratives tend to follow predictable patterns and lack diversity in tone or creativity.
* **Contextual Misalignment**: Occasionally, generated stories miss minor but crucial contextual cues (e.g., action direction or background elements).

**5.6 Inference Speed**

Inference was tested using a batch of 32 validation images:

* **Average Inference Time**: ~1.7 seconds per batch on GPU.

## **Chapter 6: Conclusions and Future Scope**

### **6.1 Conclusion**

“By utilizing artificial intelligence, particularly deep learning, this framework tackles the longstanding challenges of producing coherent and insightful narratives from static visual input. By combining inceptionv3 for extracting features from images and transformer-based models for generating sequential stories, this framework bridges the gap between image captioning and dynamic storytelling. It showcases a significant effort to replicate visual semantics and align them with consistent textual constructs, which has been a challenging but crucial task in the field of artificial intelligence-driven creative automation.

The model was successfully implemented on the flickr8k dataset and produced sensible outputs based on both the relevance of the story and its syntactic coherence. The bleu scores and user evaluations confirmed that while the model performs well for simple image inputs, it struggles with more complex or abstract scenarios. The design approach prioritizes user-friendliness, ensuring that the generated narratives are easily understandable and emotionally impactful—essential for educational tools, assistive technology, and media applications.

The solution introduces a flexible and scalable approach to automated content creation, capable of adapting to various contexts and producing content across different domains. It sets the stage for future studies on multimodal AI systems that can emulate human-like imagination and storytelling abilities.”

**6.1.1. Limitations**

These are some of the main restrictions encountered during the execution and assessment of the system that generates stories from images:

Incomplete meaning in intricate pictures.

• The model occasionally is unable to capture subtle interactions or abstract features in photos depicting intricate human or environmental situations

• More semantic parsing, such as external knowledge graphs and cultural signals, is required for increased narration accuracy

Reliance on dataset variety.

* Training was conducted mainly on the flickr8k dataset that is not domain-wide diverse.
* This limitation of the model restricts its applicability to specialized or out-of-distribution image types, such as medical or satellite photos.

High resource consumption.

• While coherent, the narratives tend to be dull and lack stylistic variety like humor, suspense, or emotion

• Conditional generation and style-specific fine-tuning can help create more expressive outputs

Assessment of Our Shortcomings

• Such measures as bleu, rouge, or spice tend to fall short in capturing the narrative quality or creativity of a narrative

• Human assessment is the most effective, albeit resource-intensive, indicator of quality

### **6.1.2 Contribution to the Discipline**

This research makes a significant contribution to the intersection of computer vision and natural language processing through the following:

Combining multiple modes of learning.

• Successfully combines image processing with sequence generation through a single encoder-decoder architecture, paving the way for the knowledge of visual-textual correspondence

Generating Narratives from Images.

• advances the state-of-the-art by shifting from static captioning to dynamic storytelling, a significant step toward creative AI.

User-friendly interface.

• accessible by design, the system enables users with no technical expertise to create stories from images, thereby democratizing ai-aided creativity

The Effects of Our Result

• has possible uses in special education, blind support aids, and narrative-based learning modules

Encouraging ethical utilization of artificial intelligence.

• emphasizes ethical considerations like bias, fairness, and interpretability, and provides guidelines for best practices in narrative ai systems

### **6.2 Future Scope**

In the future, the development of the system should focus on the following strategic directions to enhance capabilities and expand applications:

• Video story fusion with multimodal: expand the system from static pictures to sequential shots, allowing video-to-story generation for documentaries, surveillance, and entertainment

• Stylistic story customization: add tone, style, emotion, and audience-specific narrative formatting controls—e

• Zero-shot and few-shot learning: incorporate models such as clip and gpt-4/5 for improved generalization to new domains with little training data

• Cross-lingual and multilingual output: establish multilingual support to create culturally relevant tales in native tongues, enhancing international accessibility

• Reinforcement learning for quality tuning: execute reward-based learning that focuses on user-rated high-quality tales and aligns model generation with human preference

• Assistive technology deployment: establish speech-to-story systems for the blind or those who are visually impaired, which can tell surrounding stories in real time

• Edge-compatible deployment: quantize and prune the model for deployment on edge and mobile devices to enable real-time execution in resource-constrained environments

• Ethical storytelling and bias mitigation: narrative fairness and cultural sensitivity research will be critical for large-scale deployment, particularly in public or educational environments

**References**:

1. 1.O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156-3164. doi: 10.1109/CVPR.2015.7298935.

1. A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664-676, 2015. doi: 10.1109/TPAMI.2016.2598339.

1. K. Xu, J. Ba, R. Kiros, et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 2048-2057.

1. M. D. Hossain, et al., "Comprehensive Review on Image Captioning Models," *IEEE Access*, vol. 7, pp. 69564-69581, 2019. doi: 10.1109/ACCESS.2019.2932774.
2. T. Mikolov, et al., "Efficient Estimation of Word Representations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013, pp. 1-12.

1. J. Donahue, et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625-2634.

1. A. Vaswani, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998-6008.
2. S. J. Rennie, et al., "Self-Critical Sequence Training for Image Captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1179-1195. doi: 10.1109/CVPR.2017.131.

1. T. Brown, et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

1. R. Bernardi, et al., "Automatic Description Generation From Images: A Survey of Models, Datasets, and Evaluation Measures," *Journal of Artificial Intelligence Research (JAIR)*, vol. 55, pp. 409-442, 2016.

1. A. Radford, et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.
2. T. Lin, et al., "Microsoft COCO: Common Objects in Context," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740-755.
3. J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4565-4574. doi: 10.1109/CVPR.2016.494.
4. S. Jain, et al., "Incorporating Context in Image Captioning Models Using Scene Graphs," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2633-2641. doi: 10.1109/ICCV.2019.00272.
5. T. Lin, et al., "Microsoft COCO: Common Objects in Context," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740-755.

1. Sutskever, I. "Sequence to Sequence Learning with Neural Networks." *arXiv preprint arXiv:1409.3215* (2014).
2. Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).

1. Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
2. Anderson, Peter, Stephen Gould, and Mark Johnson. "Partially-supervised image captioning." *Advances in Neural Information Processing Systems* 31 (2018).
3. Dai, Andrew M., and Quoc V. Le. "Semi-supervised sequence learning." *Advances in neural information processing systems* 28 (2015).

1. Chen, Long, et al. "Human-like controllable image captioning with verb-specific semantic roles." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

1. Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer International Publishing, 2016.

1. Chen, Jun, et al. "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
2. Saleem, Mohammad Wasil, and Sandeep Uprety. "Neural Machine Translation with Attention." (2021).
3. Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383, 2021.
4. Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
5. Su, Yixuan, et al. "Language models can see: Plugging visual controls in text generation." *arXiv preprint arXiv:2205.02655* (2022).
6. Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In Proceedings of the AAAI conference on artificial intelligence (AAAI), 2016.
7. Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Typical decoding for natural language generation. arXiv preprint arXiv:2202.00666, 2022.

1. Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. CoRR, abs/2110.08173, 2021.
2. Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning. ArXiv, abs/2111.09734, 2021.
3. Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2016.
4. Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2017).

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing System