# Automated Video Narration: A Multi-Layered Approach to Captioning and Story Generation

Kartik Rodagi
*Department of Information Technology*
*NITK, Surathkal*
*kartikrodagi.211it029@nitk.edu.in*

Kushangi Sharma
*Department of Information Technology*
*NITK, Surathkal*
*kushangisharma.211it036@nitk.edu.in*

***Abstract*—This project presents a novel approach to automated video captioning and story generation by leveraging advanced deep learning and language generation models. Our system consists of two primary components: a video caption generation module and a story generation pipeline. The caption generation model is implemented using a Convolutional Neural Network (CNN) encoder-decoder architecture based on the Inception V3 model, trained on the COCO 2017 dataset, which provides a vast collection of images and their associated captions. The pretrained model is fine-tuned on selected captions for improved performance, requiring minimal epochs to achieve reliable caption outputs.**

**The story generation pipeline is designed to integrate the generated captions with user-defined thematic prompts to create coherent, engaging narratives from video content. This process begins by segmenting the video at regular intervals, capturing frames that are then fed into the caption generation model. The resulting captions, along with a customizable theme prompt, are subsequently processed by a locally hosted LLaMA 3.1 language model to generate a story that reflects the specified thematic direction. This approach allows for enhanced flexibility and novelty in storytelling, tailored to the subject and style specified by the user.**

**To evaluate the effectiveness of our system, we use short narrative videos sourced from a YouTube channel that features 60-second positive stories. Each video's original transcript and description serve as the ground truth stories, enabling a comparative analysis. The generated stories are evaluated against the original narratives using established natural language generation metrics, including BLEU, ROUGE, and other relevant scores. The results highlight the potential of combining video captioning and large language models for automated story generation, with promising applications in content creation and automated summarization.**

***Index Terms*—Video Captioning, Story Generation, Inception V3, COCO Dataset, Language Models, Evaluation Metrics, BLEU, ROUGE.**

## I. INTRODUCTION

With the exponential growth of digital content, the demand for automated systems that can interpret, summarize, and narrate multimedia content has surged. Video captioning and story generation represent key areas within this domain, aiming to enhance accessibility, comprehension, and engagement by transforming visual content into descriptive and narrative text. These technologies have applications across numerous fields, from content creation and entertainment to educational tools and assistive technologies.

In this project, we developed a comprehensive system for video captioning and story generation. The system leverages two core components: an Inception V3-based Convolutional Neural Network (CNN) model for generating captions from video frames, and a custom language model pipeline that synthesizes these captions into cohesive, theme-based stories. Our project is structured around several key objectives:

- **Automated Caption Generation**: The primary goal is to automatically generate captions for individual video frames. This is achieved by training a CNN encoder-decoder model, based on the Inception V3 architecture, using the COCO 2017 dataset, which contains diverse image-caption pairs.
- **Theme-Based Story Generation**: Building on the generated captions, our system synthesizes these into a coherent story by utilizing the LLaMA 3.1 language model. This model is locally hosted, allowing for customizable prompts that define the thematic style and direction of the generated narrative, such as romance, thriller, or positive humanistic stories.
- **Evaluation of Generated Narratives**: To ensure that our generated stories are meaningful and maintain fidelity to the video's original content, we conduct a comparative analysis. Using publicly available short story videos, we compare our generated stories to the original transcripts and descriptions, assessing accuracy and quality with evaluation metrics like BLEU and ROUGE scores.

The novelty of our approach lies in its dual-layered structure, combining frame-based video captioning with large language model (LLM)-driven story generation. By generating captions at regular intervals from a video, we establish a foundational understanding of the content, which the LLM then processes to create stories. This integration of deep learning-based visual understanding with natural language generation enables us to create dynamic, content-aware narratives that are tailored to both the video's content and the desired thematic direction.

Overall, this project contributes to the development of automated storytelling systems and illustrates the potential of advanced AI architectures to transform video into narrative form. This can enhance engagement, accessibility, and personalization in content consumption, with potential applications in various industries, including entertainment, education, and digital marketing.

## LITERATURE REVIEW

The development of video captioning and story generation systems has been significantly advanced by a range of techniques in natural language processing and computer vision. This section reviews notable approaches in image captioning and story generation that have influenced our project.

*a) Controllable Story Generation:* Peng et al. (2018) introduced methods for controlling narrative flow in AI-generated stories, enhancing coherence by using advanced natural language processing techniques [1]. This study highlighted the need for structured narrative control in story generation, though limitations were observed in the form of repetitive or formulaic outputs without further refinement.

*b) Image Captioning with Deep Bidirectional LSTMs:* Wang et al. (2021) improved image captioning by incorporating bidirectional LSTMs and multi-task learning, which enhances contextual understanding of images and produces more accurate captions [2]. However, this model is computationally intensive and may face scalability challenges with complex or cluttered images.

*c) Image-Based Storytelling:* Zhu and Yan (2020) achieved high model accuracy in image-based storytelling, achieving a mean Average Precision (mAP) of 97.21% and 99.23% [3]. Although the approach demonstrated strong performance in object detection, its reliance on pre-set templates limited the narrative complexity and impact of Transformers due to a small sample size.

*d) Systematic Review and Comparative Analysis of Storytelling Methods:* A comprehensive review by Bernardi et al. (2016) provided an analysis of models, datasets, and evaluation metrics within image captioning, offering valuable insights into the evolution of the field [4]. While highly informative, this survey did not propose new models, and newer advancements in captioning were not covered.

*e) Deep Visual-Semantic Alignments:* Karpathy and Fei-Fei (2015) introduced an alignment model that maps images and their captions into a common feature space, forming a baseline for subsequent captioning models [5]. Although effective, this approach struggled with highly descriptive captions, particularly in complex or multi-object scenes.

*f) Mind's Eye Recurrent Visual Representation:* Chen and Zitnick (2015) proposed a recurrent visual representation model to enhance temporal coherence, improving alignment between visual and textual sequences in generated captions [6]. Despite its success, this model requires substantial training data, and captions may tend toward simplicity.

*g) Semantic Attention in Image Captioning:* You et al. (2016) introduced semantic attention mechanisms, allowing the model to focus on essential areas of the image for improved contextual relevance in captions [7]. While effective, the model's complexity leads to longer training times and may require fine-tuning for diverse datasets.

*h) Microsoft COCO Dataset:* Chen et al. (2023) introduced the Microsoft COCO dataset, an influential large-scale dataset for image captioning research, providing standardized evaluation metrics to benchmark models consistently [8]. However, dataset biases and evaluation limitations are challenges to fully capturing caption nuances.

*i) Im2Text: Early Image Captioning Work:* Ordonez et al. (2011) utilized a dataset of 1 million captioned images, setting an early foundation for image captioning by leveraging large datasets to improve quality [9]. Yet, generalization remains a challenge, as the dataset's quality and diversity impact caption generation effectiveness.

The literature survey reveals significant advancements in both image captioning and story generation. While earlier models laid foundational techniques such as deep visual-semantic alignments and recurrent visual representation [5], [6], more recent studies have introduced sophisticated methods like bidirectional LSTMs and semantic attention, which enhance the contextual relevance and coherence of captions and narratives [2], [7]. However, challenges persist in scalability, generalization, and generating complex, non-formulaic stories. Furthermore, the limitations identified in current datasets and evaluation metrics indicate a need for more refined and diverse datasets, as well as new performance metrics to better assess model efficacy in real-world storytelling applications [4], [8].

## II. DATASET DESCRIPTION

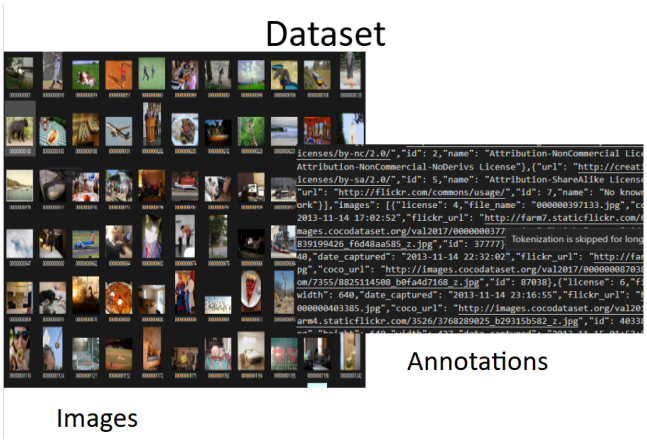


Fig. 1. Overview of MS-COCO dataset

The datasets used in this project are integral to training, evaluating, and validating the video captioning and story generation models. We utilized two primary datasets to fulfill these tasks: the COCO 2017 dataset for caption generation model training and a custom-curated set of short video stories for evaluating the system's performance.

### *COCO 2017 Dataset for Caption Generation*

The COCO (Common Objects in Context) 2017 dataset is a large-scale dataset designed to support various computer vision tasks, including object detection, segmentation, and image captioning. For our project, we specifically leveraged its extensive captioned images to train our CNN-based caption generation model. The COCO 2017 dataset provides:

- **Over 120,000 images** with detailed annotations, including five human-generated captions per image.

- **Diverse content** with images capturing complex scenes containing multiple objects, various activities, and backgrounds, enhancing the model's ability to generate rich, context-aware captions.
- **Multimodal annotation structure** that facilitates training on diverse, real-world contexts, making the model's output adaptable to different types of video frames.

By training the Inception V3-based encoder-decoder model on this dataset, we ensured that the captioning model could generate relevant and descriptive captions for a wide range of scenes.

*Custom Video Stories Dataset for Evaluation*

To evaluate the effectiveness of our video captioning and story generation pipeline, we curated a dataset of short videos from the YouTube channel "60 Seconds," known for producing concise, human-centric positive stories within a one-minute duration. This dataset includes:

- **Eight videos** with lengths of approximately 60 seconds, covering various themes in line with our story generation themes, such as positivity, human connection, and motivation.
- **Original story text files**, including video descriptions and transcripts as reference stories for each video. These text files serve as ground truth data for evaluating the accuracy and coherence of the generated stories.

The videos were processed by extracting frames at regular intervals, which were then fed into the trained captioning model. These generated captions were subsequently compiled and used as input to a locally hosted LLaMA 3.1 language model, which generated stories aligned with specified themes. The original story text files provide a benchmark for comparing the generated story's fidelity and quality using evaluation metrics like BLEU and ROUGE scores.

Together, the COCO 2017 dataset and the curated video stories dataset enable a comprehensive evaluation of both the caption generation and story synthesis capabilities of our system.

## III. Methodology

The methodology for this project comprises two core components: the Caption Generation Model and the Story Generation Pipeline. Together, these systems enable the automatic creation of descriptive captions from video frames, which are then transformed into cohesive stories aligned with specified themes.

*Caption Generation Model*

The first step in our video storytelling process involves generating captions from video frames. For this purpose, we employ a CNN-based encoder-decoder model architecture, specifically utilizing the Inception V3 model for feature extraction. The methodology for the caption generation model is outlined as follows:

- **Data Preparation:** We trained our caption generation model on the COCO 2017 dataset, which contains over 120,000 images with human-annotated captions. This dataset provides diverse contextual information, allowing the model to learn varied scene representations.
- **Model Architecture:** The caption generation model employs a CNN-LSTM encoder-decoder architecture. The CNN component, based on Inception V3, extracts high-level features from input images, serving as the encoder. These extracted features are passed to an LSTM network, which acts as the decoder, generating captions word by word.
- **Training Process:** The model was fine-tuned on the COCO dataset, leveraging the pretrained Inception V3 weights for the encoder. We trained the model for a limited number of epochs due to its pretrained state, thus ensuring efficient learning without extensive computational overhead.
- **Caption Generation for Video Frames:** After training, the model is used to generate captions for individual frames extracted from the input videos. Frames are selected at regular intervals to ensure comprehensive coverage of the video's content.

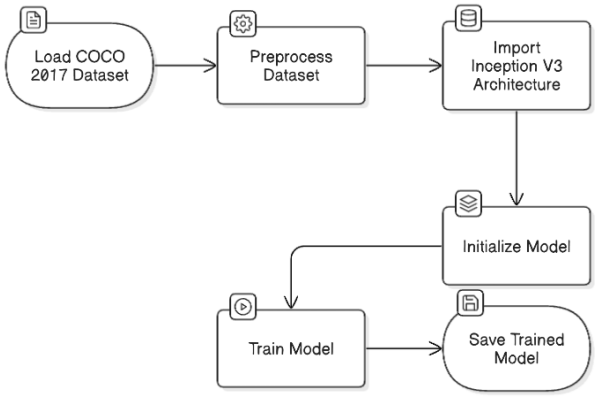


Fig. 2. Caption Generation Model

This caption generation model provides descriptive text for each frame, which forms the basis for the next step in our pipeline—the story generation process.

*Story Generation Pipeline*

Once the video captions have been generated, the next step is to transform these isolated descriptions into a cohesive story. For this, we use a story generation pipeline powered by a locally hosted large language model (LLM), specifically LLaMA 3.1, which has been fine-tuned to generate narratives based on input prompts. The process includes:

- **Frame Extraction:** Frames are extracted from each video at specified time intervals to capture the progression of the visual content. The chosen interval balances information density with processing efficiency.
- **Caption Aggregation:** Captions generated for each selected frame are aggregated into a chronological se-

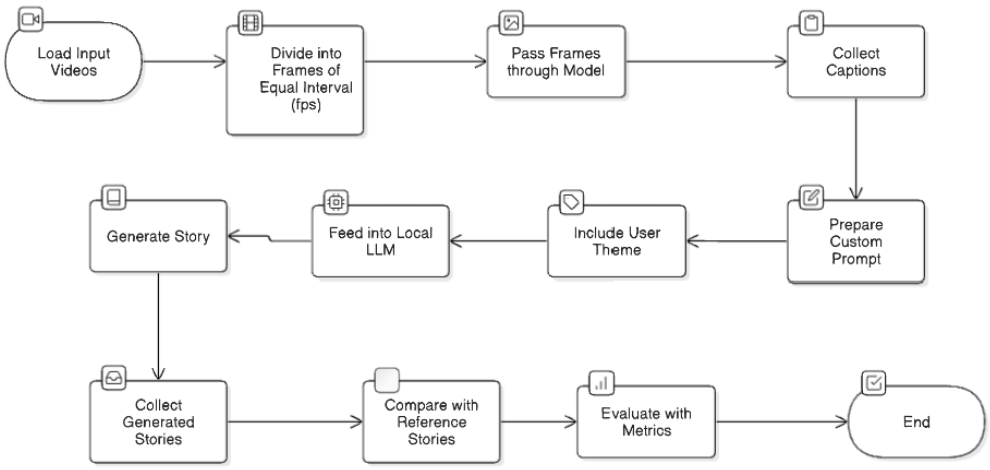


Fig. 3. Story Generation Pipeline

quence, providing a comprehensive textual overview of the video's visual content.

- **Custom Prompt Creation:** A custom prompt is generated for the LLM to guide the narrative synthesis. The prompt includes:
  - The aggregated captions from the video frames.
  - A specified theme, such as "positive humanlike story," "romantic narrative," or "inspirational theme," which directs the tone and structure of the generated story.
- **Story Generation with LLaMA 3.1:** The LLM processes the custom prompt and synthesizes a cohesive story that aligns with the theme and integrates the visual context captured in the captions. Running the model locally allows for greater control over the customization and privacy of the generation process.
- **Evaluation of Generated Stories:** The generated stories are compared to original ground-truth stories provided by the video sources. This evaluation process employs metrics such as BLEU and ROUGE scores to assess the quality, relevance, and coherence of the generated text relative to the original.

*Pipeline Summary*

Overall, this methodology combines CNN-based feature extraction, LSTM-based sequence generation, and LLM-based story synthesis to automate the process of converting video content into narrative form. By leveraging a high-quality caption generation model and a customized story generation pipeline, the system can produce meaningful, theme-aligned stories from video input.

## Result Analysis and Comparison

To evaluate the effectiveness of our video captioning and story generation pipeline, we analyzed the generated stories using several standard text evaluation metrics: BLEU, ROUGE-1, ROUGE-2, and ROUGE-L. These metrics provide insights into the similarity between the generated stories and the ground-truth stories, capturing both word overlap and sequence similarity.

*Quantitative Evaluation*

The table below presents the scores for each video, with metrics computed as follows:

- **BLEU:** Measures the precision of n-gram overlap between the generated and reference stories.
- **ROUGE-1:** Measures the overlap of unigrams (single words) between generated and reference stories.
- **ROUGE-2:** Measures the overlap of bigrams (two-word sequences) to capture short-phrase similarity.
- **ROUGE-L:** Focuses on the longest common subsequence, capturing the structure of the sentences.

| Video | BLEU | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|
| video_01.mp4 | 4.450019e-79 | 0.242697 | 0.022523 | 0.110112 |
| video_02.mp4 | 2.028977e-155 | 0.207059 | 0.014184 | 0.131765 |
| video_03.mp4 | 2.830454e-155 | 0.268398 | 0.017391 | 0.125541 |
| video_04.mp4 | 5.872773e-79 | 0.226471 | 0.029499 | 0.097059 |
| video_05.mp4 | 3.146367e-155 | 0.241486 | 0.012461 | 0.092879 |
| video_06.mp4 | 2.634403e-155 | 0.203139 | 0.020352 | 0.105263 |
| video_07.mp4 | 2.086313e-155 | 0.178071 | 0.016515 | 0.084089 |
| video_08.mp4 | 1.958548e-155 | 0.146504 | 0.006674 | 0.082131 |

TABLE I
Evaluation Scores for Generated Stories by Video

*Analysis of Results*

The results indicate varying levels of performance across different videos. Key observations from the quantitative evaluation are as follows:

- **BLEU Scores:** The BLEU scores across videos are relatively low, with values close to zero. This low BLEU performance suggests that the generated stories often diverge in word choices from the reference stories. This

divergence may be due to the abstract and creative nature of storytelling, where exact word overlap is less important than thematic coherence.

- **ROUGE-1 and ROUGE-2 Scores:** The ROUGE-1 scores range from approximately 0.14 to 0.27, indicating that the unigram overlap between the generated and reference stories is moderate. Video 03 achieved the highest ROUGE-1 score (0.268), showing that this particular story preserved more key words from the reference story. The ROUGE-2 scores are lower overall, with values ranging from 0.006 to 0.029, which suggests limited overlap in bigram sequences. This finding reflects the creative flexibility in phrase usage and sequence, which aligns with the objective of generating diverse narratives.
- **ROUGE-L Scores:** The ROUGE-L scores are moderate and indicate the longest common subsequence matches between generated and reference stories. Video 02, with a ROUGE-L score of 0.131, achieved the best structural alignment with the reference story. This suggests that while individual word choices may differ, some generated stories retain a structure that is close to the reference text.
- **Overall Performance Trends:** Video 03 consistently achieved higher scores across all metrics, showing that the generated story for this video aligns more closely with the reference text than for other videos. Conversely, Video 08 shows the lowest performance across most metrics, possibly indicating a theme mismatch or challenges in capturing the narrative style of the reference text.

*Conclusion*

The evaluation metrics reveal that while the generated stories align partially with the reference stories, there is room for improvement in achieving higher word and phrase overlap. The moderate ROUGE-1 and ROUGE-L scores indicate that the model is effective at capturing some key words and maintaining some structural similarity, though BLEU and ROUGE-2 scores suggest that sequence fidelity and exact word matching are less consistent.

Future improvements could involve fine-tuning the LLM model with additional storytelling data to enhance context sensitivity, particularly for maintaining thematic coherence in varied narrative styles. Overall, this pipeline provides a promising framework for generating stories from videos, particularly in settings where theme-oriented creativity is prioritized over exact textual match.

## Conclusion and Future Work

*Conclusion*

This project introduced a comprehensive pipeline for generating stories from videos by combining video captioning with a local LLM-based storytelling approach. The pipeline successfully captures frames from videos, generates captions using a CNN encoder-decoder model, and synthesizes these captions into cohesive stories through a story generation model. Evaluation of the generated stories against reference texts using BLEU, ROUGE-1, ROUGE-2, and ROUGE-L metrics demonstrated the model's capability to maintain thematic alignment and structural coherence, although word and phrase overlap were sometimes limited. Overall, the system offers a novel method for transforming visual content into creative textual narratives, with promising applications in content generation, automated storytelling, and entertainment.

*Future Work*

Several areas for future development have been identified to enhance the effectiveness and adaptability of this video-to-story generation pipeline:

- **Fine-tuning the Story Generation Model:** Currently, the model is reliant on a general-purpose LLM. Training or fine-tuning the model on a storytelling-specific dataset could improve its capacity to capture narrative flow and contextual depth, especially for complex or nuanced themes.
- **Improving Captioning Accuracy:** Although the inception-based captioning model performed well, incorporating an ensemble of vision-language models or leveraging the latest advancements in image-captioning models could improve the descriptive accuracy and richness of the generated captions.
- **Thematic Consistency and Customization:** Enhancing the story generation model to recognize and maintain a specified theme throughout the story could produce more cohesive narratives. Adding mechanisms to adjust for tone, character depth, or plot development based on user preferences would increase the pipeline's flexibility for varied applications.
- **Real-Time Processing and Optimization:** Implementing optimizations to accelerate the processing pipeline, including reducing LLM inference latency, would make the model more viable for real-time applications. This would also allow for broader deployment in dynamic environments, such as live event captioning or interactive media.
- **Advanced Evaluation Techniques:** The current evaluation relies on BLEU and ROUGE scores, which may not fully capture narrative quality. Future work could incorporate human evaluation, narrative coherence measures, or metrics specific to story quality and creativity to provide a more holistic assessment of the generated stories.
- **Broadening the Dataset Scope:** Expanding the dataset to include more diverse video genres (e.g., drama, action, educational content) would allow for more generalized story generation, increasing the pipeline's utility across different media formats and audiences.

By addressing these areas, the video captioning and story generation system could be refined into a robust and adaptable storytelling tool with enhanced creative control, quality, and applicability across various domains.

## References

[1] N. Peng, M. Ghazvininejad, J. May, K. Knight. Towards controllable story generation. 2018.

[2] C. Wang, H. Yang, C. Meinel. Image captioning with deep bidirectional LSTMs and multi-task learning. 2021.
[3] Y. Zhu, W. Qi Yan. Image-Based Storytelling Using Deep Learning. Auckland University of Technology, New Zealand. 2020.
[4] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. 2016.
[5] A. Karpathy, L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. 2015.
[6] X. Chen, C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. 2015.
[7] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo. Image captioning with semantic attention. 2016.
[8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. 2023.
[9] V. Ordonez, G. Kulkarni, T. L. Berg. Im2text: Describing images using 1 million captioned photographs. 2011.