Image Caption Generator

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of CSE**

**Deep Learning with 23AVI3101A**

Submitted by

2310030003 Bhavani Tadimalla

Under the guidance of

**Dr. Sumit Hazra**



Department of Computer Science and Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

August- 2025.

**Introduction**

Image caption generation is a challenging problem at the intersection of computer vision and natural language processing (NLP). The goal is to automatically generate a meaningful description of a given image in natural language. For humans, describing an image is effortless, but for machines it requires both understanding the visual content and expressing it in grammatically correct sentences. This makes the task a perfect example of combining perception and language understanding.

The process typically involves two main components: an encoder and a decoder. The encoder, usually a Convolutional Neural Network (CNN) or a Vision Transformer, extracts high-level feature representations from the image. The decoder, often a Recurrent Neural Network (RNN) such as Long Short-Term Memory (LSTM) or a Transformer-based language model, takes these features and generates captions word by word. Recent advancements also integrate attention mechanisms, allowing the model to focus on specific regions of the image while producing each word.

Image caption generators have practical applications in assistive technologies for the visually impaired, content management, image retrieval, autonomous systems, and smart devices. With growing datasets and deep learning innovations, research continues to improve caption quality, handling ambiguity, unseen objects, and generating more descriptive and context-aware captions.

**Literature Review/** **Application Survey**

Image caption generation has emerged as one of the most significant tasks in the field of artificial intelligence, combining the domains of computer vision and natural language processing (NLP). The objective is to automatically generate human-like descriptions of images. This problem is complex because it requires a system to not only recognize the objects present in an image but also understand their relationships, context, and semantics before expressing them in natural language. Over the past decade, extensive research has been carried out in this area, leading to several landmark models and applications.

**Literature Review**

**Early CNN-RNN Based Models**

The first successful attempts at image captioning relied on the integration of Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for sequence generation. Google’s “Show and Tell” model (2015) was one of the earliest frameworks that used a pre-trained CNN (Inception) to encode image features and a Long Short-Term Memory (LSTM) network as the decoder. The CNN was responsible for capturing visual semantics, while the LSTM generated sentences word by word. This model was trained end-to-end on large datasets such as MS-COCO and set the foundation for many later studies.

**Attention Mechanisms**

While the Show and Tell model was a breakthrough, it treated the image as a single fixed feature vector, which limited its ability to describe specific details. To address this, the “Show, Attend and Tell” model (2015/16) introduced the use of attention mechanisms. Instead of processing the entire image equally at every time step, attention allowed the decoder to focus on different spatial regions of the image when generating each word. This improvement enabled captions to be more context-specific and detailed, significantly enhancing caption accuracy.

**Object-Level Features**

Later works such as Bottom-Up and Top-Down Attention (2018) refined the attention mechanism further by incorporating object detection models like Faster R-CNN. Instead of only using global CNN features, these models extracted region-level features corresponding to individual objects and applied attention to them. This approach allowed captions to highlight fine-grained details, such as “a young boy holding a red baseball bat,” rather than generic phrases like “a boy playing outside.”

**Reinforcement Learning and Evaluation Metrics**

One limitation of supervised training with cross-entropy loss is that it does not directly optimize for the metrics used to evaluate captions, such as BLEU, METEOR, ROUGE, CIDEr, and SPICE. To solve this, Rennie et al. proposed Self-Critical Sequence Training (SCST, 2017), which framed caption generation as a reinforcement learning problem. The model was rewarded based on its CIDEr score, enabling it to produce captions that aligned better with human-judged quality. This marked a significant shift in how captioning systems were optimized.

**Transformer-based Approaches**

With the rise of Transformers in NLP, image captioning also transitioned to self-attention based architectures. Unlike RNNs, Transformers process sequences in parallel and can capture long-range dependencies more effectively. Models such as Meshed-Memory Transformer (2019) and AoANet (Attention on Attention, 2020) improved caption fluency and coherence. Transformers allowed for more flexible integration of image and language features, reducing the reliance on sequential processing.

**Vision-Language Pretraining**

The most recent advancements in image captioning come from large-scale vision-language pretraining (VLP) models. Architectures like OSCAR, VinVL, BLIP, BLIP-2, OFA, and GIT are trained on millions of image-text pairs scraped from the internet. These models learn joint embeddings of images and text, enabling them to perform captioning with much higher accuracy and even work in zero-shot or few-shot scenarios. Another popular model, CLIP (Contrastive Language-Image Pretraining, OpenAI 2021), learns aligned embeddings for text and images, which has been used as a backbone for modern captioning pipelines.

**Datasets**

Image captioning research has been fueled by several benchmark datasets:

**Flickr8k/Flickr30k –** Early datasets with thousands of captioned images.

**MS-COCO –** The most widely used dataset, containing over 120,000 images with five captions each.

**SBU Captions & Conceptual Captions –** Large-scale web-scraped datasets for pretraining.

**no caps –** A dataset specifically designed to test the ability of models to handle unseen objects.

**Application Survey**

The development of image captioning models has opened up numerous real-world applications:

**Accessibility for the Visually Impaired**

One of the most impactful applications is in assistive technologies, where captioning models can describe images to visually impaired users. Integrated into screen readers or wearable devices, these models enable independent navigation and understanding of visual content.

**Search and Indexing**

Digital platforms often host billions of images. Automatic captioning helps in content indexing and image retrieval by generating searchable metadata. For example, social media platforms can auto-generate captions for uploaded photos, improving discoverability.

**E-commerce Applications**

Online shopping platforms benefit from captioning systems that describe product images automatically. This enhances product descriptions, improves recommendations, and boosts customer engagement.

**Social Media and Content Creation**

Caption generators are used to create auto-captions, tags, and hashtags for user-generated content. They also support content moderation by identifying inappropriate or harmful imagery.

**Robotics and Autonomous Vehicles**

Robots and self-driving cars can use captioning to provide scene understanding. For instance, a robot could summarize its environment (“a person standing in front of a door”) to interact better with humans.

**Medical Imaging**

In healthcare, captioning can assist radiologists by automatically generating preliminary descriptions of X-rays, CT scans, and MRI images. With fine-tuning, captioning models can be adapted for domain-specific tasks.

**Education and Creative Tools**

In educational contexts, captioning can enrich teaching materials, provide image-based quiz descriptions, and support language learning. Creative software integrates caption generators for storytelling, photo albums, and smart photo apps.

**Summary**

The research trajectory in the field of image caption generation clearly reflects a continuous effort to bridge the gap between visual understanding and natural language expression. Early works were primarily based on the CNN-LSTM paradigm, where Convolutional Neural Networks (CNNs) were used as powerful feature extractors and Long Short-Term Memory (LSTM) networks served as decoders to generate sentences word by word. While these models achieved initial success, they often produced generic captions and struggled to capture fine-grained details within complex scenes.

To overcome these shortcomings, researchers introduced attention mechanisms, marking a significant turning point in captioning research. Attention allowed the model to dynamically focus on different regions of the image when generating each word, thereby enabling more descriptive and context-aware captions. For example, instead of merely saying “a man with a dog,” attention-based systems could refine the description to “a man walking his brown dog on a grassy field,” demonstrating the ability to highlight specific attributes and relationships in the image. This development brought image captioning models closer to human-like descriptive ability.

Despite the advances made with attention, a major limitation remained in the optimization process. Most models were trained using cross-entropy loss, which does not directly align with evaluation metrics like BLEU, METEOR, CIDEr, or SPICE. This gap motivated the adoption of reinforcement learning (RL)-based approaches, particularly Self-Critical Sequence Training (SCST). By framing caption generation as a sequential decision-making task, RL methods enabled models to directly optimize for caption quality as measured by evaluation metrics. This shift led to captions that were not only syntactically correct but also semantically richer and more aligned with human evaluation.

The next major leap came with the adoption of Transformer-based architectures. Transformers, originally introduced in NLP, revolutionized captioning models by eliminating the limitations of sequential processing inherent in RNNs. Unlike LSTMs, Transformers employ self-attention to process entire sequences in parallel, enabling them to model long-range dependencies more effectively. This architecture provided superior fluency, coherence, and scalability, making it more adaptable for large datasets. Moreover, advanced models such as Meshed-Memory Transformer and AoANet demonstrated that integrating multi-head attention into image captioning could further improve semantic understanding and diversity in generated captions.

In recent years, the field has been transformed by large-scale vision-language pretraining (VLP). Models such as OSCAR, VinVL, BLIP, BLIP-2, and GIT leverage massive collections of image-text pairs from the web to learn joint embeddings of images and language. These models are capable of zero-shot and few-shot captioning, where captions can be generated without explicit supervised training on a given dataset. Additionally, multimodal models like CLIP have shown that vision-language alignment at scale enables strong generalization, making captioning systems more robust and versatile than ever before. This era marks the shift from task-specific architectures to generalized multimodal AI systems.

The applications of image captioning models span a wide spectrum, reflecting their societal and industrial relevance. In assistive technologies, these systems have been integrated into tools for the visually impaired, providing spoken descriptions of images in real-time and enhancing independence. In e-commerce, automated captioning assists in product cataloging, search optimization, and personalized recommendations. Social media platforms employ captioning for automatic tagging, content indexing, and moderation, thereby improving user engagement and safety. In autonomous systems and robotics, captioning supports situational awareness by enabling machines to describe their environments. Additionally, healthcare applications show promise, where medical images such as X-rays or MRIs can be automatically described, supporting doctors with preliminary observations and reducing diagnostic workload.

In summary, the literature reflects a clear progression of methodologies: from CNN-LSTM based models, to attention-enhanced architectures, to reinforcement learning optimization, and finally to Transformer-based and pre-trained vision-language models. Each generation has incrementally addressed the weaknesses of its predecessors, producing captions that are more accurate, detailed, and context-aware. The continued development of multimodal models signals a promising future, where image captioning systems are expected to become more robust, human-like, and adaptable across diverse domains. With growing datasets and advancements in deep learning, image captioning stands as a critical research area, bridging vision and language while driving innovation in accessibility, automation, healthcare, and beyond.