IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 5, December 2025

Impact Factor: 7.67

A Comprehensive Survey on Deep Learning-Based Architectural Image Captioning for Visual Accessibility

Hima R, Sangeetha VL, Rachana Ashok, Dr. Kavita Patil, Rakshitha S
Department of Information Science and Engineering
Global Academy of Technology, Bengaluru, India
himar1ga22is063@gmail.com, sangeetha1ga22is136@gmail.com, rachana1ga22is117@gmail.com
kavitapatil@gat.ac.in, rakshitha1ga22is124@gmail.com

Abstract: *A challenging issue that is receiving more and more attention in the field of artificial intelligence is image captioning. Among other applications, it can be used for intelligent blind guidance, human-computer interaction, and effective picture retrieval. This article examines developments in deep learning-based image captioning techniques, such as the encoder-decoder structure, enhanced encoder and decoder techniques, and other enhancements. We also talk about potential avenues for further research.*

Keywords*:* Artificial Intelligence

I. INTRODUCTION

Because there are so many unlabelled photographs on the internet, manual labelling is not feasible. With applications in effective image retrieval, intelligent aid for the visually impaired, and human-computer interaction, automatically producing natural language descriptions for images—also referred to as image captioning—is a valuable and difficult task in artificial intelligence..

Accurate and insightful descriptions of the provided images are the aim of image captioning. Accurately identifying objects, properties, semantic linkages, and location information is necessary for this. Consequently, picture captioning may be broken down into two primary subtasks: (1) accurately acquiring visual information through image understanding, and (2) creating descriptions based on that knowledge. Natural language processing (NLP) and computer vision (CV), two important areas of artificial intelligence, are connected by this problem.

Traditional feature extraction techniques used manually created operators to record colour, texture, and geometry, which were then merged to create high-level features. However, these approaches were limited by the "semantic gap," which occurs when low-level features are unable to convey complicated semantics, and their need on human expertise. As a result, conventional methods were neither generalisable or robust. The two approaches used by earlier models were template-based and retrieval-based. While template-based methods identify visual features and incorporate them into organised templates, retrieval-based systems choose captions from preset sets of image descriptions. Nevertheless, template-based captions lack diversity and flexibility, while retrieval-based captions frequently do not appropriately match the image content.

This discipline underwent a revolution with the advent of deep learning. While recurrent neural networks (RNNs) became essential for natural language processing, convolutional neural networks (CNNs) shown remarkable effectiveness in visual tasks including object detection and image categorisation. Deep learning-based captioning systems began when Vinyals et al. (2015) proposed an image captioning model that uses GoogleNet as the encoder to extract image features and long-term memory (LTSM) as the decoder to generate descriptive sentences. This model was inspired by the encoder-decoder framework of machine translation (Sutskever et al., 2014).

Many studies have since improved this framework. Enhancements have concentrated on improving the model's overall performance, the encoder's visual representation, and the decoder's ability to generate more coherent language.

DOI: 10.48175/IJARSCT-30670

Semantic attention (You et al., 2016), visual sentinel (Lu et al., 2017), and revision networks (Yang et al., 2016) are notable developments that improve contextual fluency and visual-semantic alignment.

This article's primary contributions are: (1) a review of the encoder-decoder framework; (2) an analysis of conventional retrieval- and template-based approaches; (3) an overview of advancements made to the encoder and decoder components; and (4) a suggestion for future research areas.

This is how the article is structured: Traditional image captioning techniques are covered in Section 2, encoder-decoder enhancements are covered in Section 3, datasets and evaluation metrics are described in Sections 4 and 5, future research possibilities are highlighted in Section 6, and the study on the total number of words is concluded in Section 7.

II. RELATED WORKS

The Encoder-Decoder model was widely used in early picture captioning research, setting the stage for deep learning-based techniques. Setting the standard for later work, Vinyals et al. (2015) presented Show and Tell, a groundbreaking model that used an LSTM as the decoder and a CNN (GoogleNet) as the encoder. By utilising Soft and Hard Attention methods, Xu et al. (2015) improved this design, enabling the model to concentrate on the image's most prominent areas and increase captioning accuracy.

Yang et al. (2016) created Revision Networks to improve feature abstraction during decoding, whereas You et al. (2016) suggested Semantic Attention to combine visual and semantic characteristics, producing more intelligible captions. Fu et al. (2017) suggested a region-based attention method to generate scene-aware captions, whereas Chen et al. (2017) presented SCA-CNN, which combines spatial and channel-wise attention for richer contextual comprehension. In order to enhance spatial feature modelling, Liu et al. (2017) developed the Multimodal Attentive Translator (MAT), which combines object detection and attention. Visual Sentinel Gates were used in further research, such as Lu et al. (2017), to improve accuracy on non-visual words through adaptive attention. Text-Conditional Attention was first presented by Zhou et al. (2017), who focused on linguistic context during generation.

Recent contributions include Anderson et al. (2018), whose Bottom-Up and Top-Down Attention produced state-of-the-art results by combining Faster R-CNN with LSTM-based decoding, and Yao et al. (2018), who modelled object interactions using Graph Convolutional Networks (GCNs). Subsequent models examined architectural efficiency: Aneja et al. (2018) used CNNs as decoders, Wang and Chan (2018) offered CNN+CNN structures for parallel processing, and Dai et al. (2018) deployed GRU decoders to preserve spatial characteristics.

Overall, these developments steadily increased semantic correctness, contextual awareness, and computational efficiency, demonstrating a clear transition from static encoder-decoder designs to dynamic, attention-centric captioning architectures.

III. LITERATURE REVIEW

From traditional, hand-crafted methods to complex deep learning-based models that combine attention, spatial awareness, and semantic reasoning, the area of image captioning has swiftly advanced. The Encoder-Decoder architecture served as the foundation for many early investigations and was later extended to a variety of network designs intended to enhance linguistic fluency and visual comprehension. The main contributions from 2015 to 2018 are methodically reviewed in the part that follows, with an emphasis on their approaches, innovations, benefits, and drawbacks.

A. Earlier Architectures for Encoder-Decoder

Vinyals et al. (2015)'s groundbreaking study Show and Tell: A Neural Image Caption Generator signalled a paradigm shift in captioning images automatically. The authors used an encoder-decoder architecture in which a Long Short-Term Memory (LSTM) network produced English language words and a Convolutional Neural Network (CNN), especially GoogleNet, extracted visual data. This model showed how deep learning may successfully connect language and visual representations. It served as a standard for subsequent advancements due to its robustness and simplicity. But because it lacked an attention mechanism, it was unable to concentrate on areas of the image with fine texture.

DOI: 10.48175/IJARSCT-30670

By adding soft and hard attention methods to the Encoder-Decoder framework, Xu et al. (2015) developed Show, Attend, and Tell in order to get around this restriction. The model improved its accuracy and interpretability by dynamically focussing on areas of the image that were considered significant. Notwithstanding these developments, the architecture's processing demands resulted in lengthier training periods and increased memory use.

B. The Development of Contextual and Semantic Attention

With Semantic Attention Image Captioning, You et al. (2016) improved the field by combining visual information with semantic descriptors taken from CNNs. This hybrid attention approach helped create semantically richer captions and enhanced contextual awareness. Nevertheless, attribute detection necessitated more annotated data and complicated model training.

Yang et al. (2016) included LSTM decoders, attention layers, and revision modules in their proposal for Revision Networks for Caption Generation that same year. This framework enhanced feature abstraction and linguistic coherence by enabling the model to recode intermediate features prior to sentence creation. Revision layers considerably raised model complexity and training costs even if performance improved.

SCA-CNN, or Spatial Attention and Channel-by-Channel CNN, was introduced by Chen et al. (2017) and enabled the model to simultaneously capture dependencies between spatial and feature channels. Higher captioning accuracy and enhanced visual reasoning were attained by the dual-attention framework. However, because attention was dispersed across dimensions, it necessitated significant computational resources and lengthy training time.

C. Scene- and Region-Aware Captioning

In their 2017 work, Aligning Where to See and What to Tell, Fu et al. combined scene context based on Latent Dirichlet Allocation (LDA) with region-based attention. More organic and contextually aware captions were produced as a result of the system's alignment of visual regions with matching descriptive elements. The primary drawback of the strategy was its reliance on the region proposal's quality; subpar suggestions resulted in lower performance.

The Multimodal Attentive Translator (MAT), created by Liu et al. (2017), also used a Seq2Seq model that included object detection and attentiveness. Through the integration of multimodal data, the method successfully modelled spatial characteristics. Longer training periods and more thorough preprocessing were necessary, nevertheless.

The Visual Sentinel Model, as out by Lu et al. (2017), incorporates an adaptive attention mechanism that chooses whether to use language context or visual cues. Due to selection methods, this invention increased computing overhead but also enhanced accuracy, particularly for non-visual terms.

Conditional attention to text, in which attention weights were conditioned on previously created words, was incorporated into the design of Watch What You Just Said by Zhou et al. (2017). Though it was prone to overfitting on small datasets, this approach enhanced contextual consistency between generated phrases and visual cues.

D. A Reasoning Models

eWork generated sentences using a top-down LSTM and extracted object-level features using Faster R-CNN. Combining top-down (language-based method) and bottom-up (region proposals) attention greatly enhanced interpretability and produced state-of-the-art (SOTA) outcomes. Slower inference because of the region proposal phase was the trade-off, though.

Using a multi-layer LSTM in conjunction with attention mechanisms, Fang et al. (2018) presented Look Deeper and Transfer Attention. The model showed enhanced verb and adjective learning, which resulted in captions with more linguistic depth. However, the computational complexity and training time increased with this depth.

E. The CNN-Based and Hybrid Architectures

Convolutional Image Captioning was introduced by Aneja et al. (2018) in response to the growing emphasis on efficiency, substituting CNN-based decoders for recurrent networks. This architecture made parallel computing possible, which sped up inference and training. Nevertheless, it periodically lost sequential context, which had an impact on the captions' grammatical fluency.

DOI: 10.48175/IJARSCT-30670

In a similar vein, Wang and Chan (2018) used CNNs for both encoding and decoding when they created CNN+CNN: Convolutional Decoders for Image Captioning. The model's language fluency slightly decreased, but it demonstrated substantial parallelisation and less dependence on sequential processing.

Last but not least, Dai et al. (2018) presented Rethinking the Shape of Latent States in Image Captioning, which preserved 2D feature maps within the decoder by substituting Gated Recurrent Units (GRUs) for LSTM units. By maintaining spatial organisation throughout the generating process, this design enhanced visual coherence.

F. The Summary and The Outlook

Multimodal, graph-based, and high-attention models replaced simple encoder-decoder architectures in image captioning research between 2015 and 2018. Every generation of models improved the semantic foundation, optimised spatial reasoning, optimised computational efficiency, and addressed specific restrictions, starting with inattention. The most significant invention turned out to be attention mechanisms, which enable networks to dynamically concentrate on pertinent words and regions, enhancing accuracy and interpretability.

Even with great advancements, difficulties still exist. Context generalisation, dataset bias, computational overhead, and real-time scalability are issues with current approaches. Additionally, whereas multimodal and graph-based methods enhanced reasoning, their general use is constrained by the architectural complexity they created. For adaptive caption optimisation, future research should focus on lightweight hybrid architectures that include reinforcement learning, cross-modal embeddings, and transformer-bas.

Conclusion: From basic visual-linguistic mapping to effective, semantically rich, and contextually aware caption generation systems, the studied literature shows a clear progression in approach and capabilities. This corpus of work establishes the foundation for next models that seek to use natural language to comprehend images in a manner similar to that of humans.

IV. COMPARATIVE ANALYSIS OF CURRENT SYSTEMS

Recent years have seen a substantial advancement in image captioning research, with techniques moving from simple encoder-decoder designs to more complex attention-based and graph-based models. Despite its limits in capturing fine-grained visual information, previous work like "Show and Tell" (Vinyals et al., 2015) used CNNs in conjunction with LSTMs for caption creation, establishing the groundwork for a model. In order to enhance spatial and semantic focus inside images, subsequent research incorporated a variety of attention processes, which raised the relevance and accuracy of captions.

To improve abstract characteristics and optimise caption semantics, recent developments have included semantic attention (You et al., 2016) and reviewer networks (Yang et al., 2016). The problem of capturing pertinent visual regions and contextual awareness is further addressed by spatial attention (Chen et al., 2017) and region-based attention that incorporates scene context (Fu et al., 2017). Multimodal inputs and object detection are used in models like MAT (Liu et al., 2017) to provide more complex captions, but frequently at the expense of higher computational overhead.

In order to improve learning efficiency, more recent models use increased attention layers and graph convolutional networks to highlight fine-grained interactions inside pictures (Yao et al., 2018; Fang et al., 2018). The balance between model efficiency, caption quality, and computational complexity is still a priority. For real-time applications, CNN-based decoders and two-dimensional representations of latent states (Aneja et al., 2018; Dai et al., 2018) strive for quicker processing and spatial structure retention.

TABLE I. Literature comparison: methods, highlights, and negatives

Reference & Authors	**Approach**	**Strengths**	**Limitations**
Vinyals and associates (2015)	Encoder-Decoder CNN+LSTM	Fundamental design; based on analogy Inattentive people miss nuances.	ignores nuances
Xu and	Encoder, Decoder, and	increases accuracy by	computationally costly

colleagues (2015)	Soft/Hard Focus	concentrating on key areas of the image.	
You and colleagues (2016)	Semantic Attention+ CNN and LSTM	blends visual and semantic elements	Enhanced intricacy as a result of attribute detection
Yang and associates (2016)	LSTM + Reviewer + Attention	Enhances the abstraction of features	A more complex model
Chen and associates (2017)	Attention Mechanism of SCA-CNN	records channel and geographical data.	More time spent training
Fu and associates (2017)	Attention depending on region plus LDA	Naturalness is improved with scene-aware captions.	Depending on how well the region proposal is written
Liu and associates (2017)	Seq2Seq + Object Detection + Attention	accurately depicts geographical features	Thorough training and preprocessing
Yao and associates (2018)	GCN + Semantic Graphs + Faster R-CNN	optimises captions by utilising object relationships.	intricate design
Anderson et al., 2018	Faster R-CNN + Downstream LSTM	Strong foundation; cutting-edge outcomes	Inference is slow because of the proposal step.
In 2018, Fang et al.	Multi-layer Attention + LSTM	Improved understanding of verbs and adjectives	More time spent training

Generalisation capability: These models are effective in few-shot and zero-shot situations and exhibit robust cross-domain generalisation thanks to pretraining language and vision on big datasets.

Multimodal Reasoning: Baseline models provide contextual and human-level verbal fluency for a variety of settings by integrating reasoning across textual and visualmodalities.

Unified Architecture: Current systems employ architectures where text and images interact through transformation layers for end-to-end learning, which improves scalability and adaptability, as opposed to modular designs (separate CNNs and RNNs).

Despite these developments, there are still significant obstacles. Significant processing power is needed for transformer-based models, particularly when training on extensive visual language datasets. For real-time applications or edge deployments, their large memory and computational requirements may act as a bottleneck, hence research on lightweight alternatives and optimisation strategies is ongoing. Additionally, although user-aligned benchmarks (like CapArena-Auto or BLIPScore) have supplemented quantitative metrics like BLEU, METEOR, and CIDEr, the development of accurate and nuanced evaluation measures is still ongoing, particularly for subjective aspects like creativity or subtitle usefulness.

Due to their deep learning flexibility, PyTorch and TensorFlow form the foundation of the majority of state-of-the-art systems in practical implementations, and training data is growing with more diverse and bigger multimodal corpora. Recent baseline models' open-source nature promotes industry adoption, innovation, and reproducibility.

Further research avenues include investigating improved cross-modal learning techniques for under-represented domains, creating transformer variations that preserve excellent subtitle quality while lowering inference latency and training expenses, and guaranteeing equity and minimising bias in generated subtitles. Additionally, new evaluation techniques that are even more in line with human judgements and their consequent utility are expected.

In summary, the development of picture captioning has progressed from the initial CNN-RNN stacks to strong visual language converters, leading to models that are more flexible, efficient, and sensitive to context. Now, the emphasis is on resource efficiency, implementation scalability, and the capacity to support a broad variety of intelligent visual language applications, in addition to descriptive accuracy.

DOI: 10.48175/IJARSCT-30670

V. RESEARCH GAPS & FUTURE DIRECTIONS OBJECTIVES

The shift from proof-of-concept studies to scalable real-world applications is hampered by a number of crucial research gaps that persist despite notable advancements in deep learning models for picture captioning, particularly in specialised fields like architecture.

1. The Dataset Limitations and Standardization

Existing architectural picture captioning datasets frequently lack diversity, scope, and annotation quality, which leads to models that are not very generalisable to other architectural settings and styles. Large-scale, standardised datasets with thorough, excellent annotations spanning a variety of architectural features, design philosophies, and contextual data are desperately needed. The model's worldwide applicability will be enhanced by producing extensive datasets with multilingual captions, guaranteeing linguistic and cultural inclusion.

2. Fusion of Integrated Multimodal Data

The majority of current models ignore complementing modalities like verbal descriptions of architectural principles or spatial data in favour of concentrating exclusively on visual elements that are extracted using CNNs. In order to create more thorough and precise captions, future studies should concentrate on multimodal data fusion techniques that integrate textual, visual, and even spatial data. More contextual and significant descriptions can be produced by combining building requirements, semantic annotation, and 3D architectural models.

3. Speech and Multilingual Interfaces

Even though multilingual translation has advanced, there is still a research gap in the creation of models that can produce captions in multiple languages at once with a high degree of accuracy and naturalness. In order to provide more inclusive architectural visualisation tools, future systems should integrate cutting-edge multilingual transformers and voice synthesis technologies that enable speech output for a variety of user groups, including professionals and users with visual impairments.

4. User-centred, Real-Time System Design

Real-time capabilities, which are necessary for interactive applications like virtual tours and on-site architectural inspections, are frequently absent from current implementations. Lightweight CNN-LSTM architectures tuned for low-latency processing, either via model pruning or quantisation, should be investigated in order to overcome this. For a smooth user experience, backend models should also be connected with graphical user interface (GUI) frameworks like Tkinter, which provide real-time audio playback, captioning, and easy image loading.

5. Adaptability and Transfer Across Domains

For models trained on small datasets, the wide regional and historical variations in architectural styles present a problem. It is important to research domain adaptation and transfer learning strategies so that models can successfully adjust to new contexts or styles with little retraining.

6. Metrics for Assessment and Comparison

Current measures, including BLEU and METEOR, evaluate correctness but fall short in capturing user happiness, creativity, and contextual relevance. A more thorough evaluation of system performance can be obtained by creating domain-specific benchmarks and adding human-informed assessments. Standardised testing procedures for different architectural situations should be established in order to enable equitable model comparisons.

7. Legal and Ethical Aspects to Consider

The growing usage of copyrighted architectural designs and personal information about building spaces raises ethical questions about consent, ownership, and data privacy. To create reliable systems, future research should incorporate privacy-preserving strategies, open data governance, and adherence to legal requirements.

DOI: 10.48175/IJARSCT-30670

8. Explainability and Ethical AI

It is crucial to make sure that models provide clear and understandable captions, particularly in the architectural field where the generated descriptions are used to inform professional judgements. In order to support model outcomes and promote user confidence, explainability procedures must be incorporated.

A multidisciplinary strategy including computer scientists, architects, ethicists, and legislators is needed to close these gaps. The development of datasets should be the main focus of future studies. scalable, inclusive, and standardised models; the creation of effective, real-time multimodal models; and the implementation of open governance and assessment procedures. Together, these initiatives will allow CNN-LSTM-based architectural captioning systems to evolve from experimental prototypes to useful instruments that improve architectural visualisation, training, and decision-making in an ethical and user-focused way.

VI. CONCLUSION

An important development in artificial intelligence, namely in the areas of accessibility and human-computer interaction, is the combination of CNNs, LSTMs, and TTSs for multimodal image captioning. Given the large number of unlabelled images available online and the impracticality of manual annotation, this work effectively tackles the difficulty of producing contextual and accurate natural language descriptions for photographs. The device gives significant advantages to visually impaired users by giving relevant audio input, proving a useful application that enhances inclusion and educational opportunities.

This work uses a novel encoding-decoding architecture to bridge the gap between computer vision and natural language processing, two core areas of artificial intelligence. The model solves the drawbacks of earlier template-based captioning or retrieval techniques as well as conventional manual captioning methods by using CNNs for reliable visual feature extraction and LSTMs for sequential language creation. The system may overcome constrained and rigid templates or uneven retrieval results by utilising the deep learning paradigm, which allows it to capture the intricate semantic linkages and spatial information required for high-quality subtitle production.

The optimised model performance is a result of notable advancements in the encoder and decoder designs. While the LSTM-based decoder produces grammatically sound and contextually relevant descriptions, the encoder effectively extracts semantic visual signals. The use of text-to-speech technology makes the output mode more accessible and strengthens the research's usefulness. This combination of multiple modalities not only makes digital content for people with visual impairments but also creates new opportunities for multimodal understanding and generation in future AI research.

This work, which reflects developments in recent literature including semantic attention approaches, visual sentinel models, and review networks, emphasises the significance of semantic attention mechanisms and contextual alignment for enhancing subtitle quality. These methods guarantee that the output language remains relevant and fluent while the model retains sensitivity to objects and scene characteristics. Combining these approaches into a single system is a significant step towards resolving the semantic gap issues that have previously hindered automated image description.

All things considered, this study offers a scalable and successful solution to the image captioning issue, with immediate ramifications for enhancing accessibility technologies and human-computer interactions. It illustrates how using deep neural networks in a multimodal context may produce precise and insightful descriptions, hence resolving important AI problems with comprehending and characterising intricate visual input. Additionally, it establishes the foundation for upcoming architectural advancements that might integrate more complex multimodal inputs and outputs, enhancing the resilience and adaptability of intelligent captioning systems.

As a result, this paper offers a thorough and useful method for multimodal image captioning that combines accessible audio feedback, contextual language modelling, and strong visual feature extraction. It provides a strong basis for future research and development in accessible AI technologies and demonstrates the revolutionary potential of deep learning to bridge language and vision.

DOI: 10.48175/IJARSCT-30670

REFERENCES

[1]. Sutskever et al. (2014) Q. Le, O. Vinyals, and I. Sutskever. Sequence learning using neural networks. (2014), NIPS, pp. 3104–3112.

[2]. Vedantam et al. (2015) R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Using consensus to evaluate photo descriptions. CVPR, 2015, pp. 4566–4575.

[3]. [In 2015, Vinyals and others] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A neural network-powered image caption generator. Page 3156–3164, CVPR, 2015.

[4]. [Xu and others, 2015] A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, J. Ba, R. Kiros, K. Cho, and K. Xu. Show, Attend, Tell: Using Visual Attention to Generate Neural Image Captions. ICML, 2015, pp. 2048–2057.

[5]. [Yang and others, 2016] Z. Yang, R. Salakhutdinov, W. W. Cohen, Y. Yuan, and Y. Wu. Examine networks for the creation of captions. NIPS, 2016, pp. 2361–2369.

[6]. [Yao and others, 2018] T. Mei, Y. Li, Y. Pan, and T. Yao. investigating visual relatedness for captioning images. Pages 711–727 in ECCV, 2018.

[7]. [You and others, 2016] J. Luo, C. Fang, Z. Wang, H. Jin, and Q. You. Image captioning with semantic attention. Pages 4651–4659 in CVPR, 2016.

[8]. [Young and others, 2014] J. Hockenmaier, M. Hodosh, A. Lai, and P. Young. New Similarity Metrics for Semantic Inference on Event Descriptions: Transitioning from Image Descriptions to Visual Denotations. TACL, 2014, 2:67–78.

[9]. [Zhou and others, 2017] J. J. Corso, P. A. Koch, C. Xu, and L. Zhou. Take note of what you just said: Text-conditional attention combined with image captioning. Pages 305–313, Proceedings of the ACM Multimedia Topical Workshops, 2017.

[10]. [Wang and Chan, 2018] A. B. Chan and Q. Wang. CNN+CNN: Convolutional Image Captioning Decoders. 2018; CoRR, abs/1805.09019.

DOI: 10.48175/IJARSCT-30670