A comprehensive survey on deep learning-based architectural image captioning for visual accessibility

Hima R
Department of Information Science and Engineering
Global Academy of Technology
Bengaluru, India
himar1ga22is063@gmail.com

Rachana Ashok
Department of Information Science and Engineering
Global Academy of Technology
Bengaluru, India
rachana1ga22is117@gmail.com

Rakshitha S
Department of Information Science and Engineering
Global Academy of Technology
Bengaluru, India
rakshitha1ga22is124@gmail.com

Sangeetha VL
Department of Information Science and Engineering
Global Academy of Technology
Bengaluru, India
sangeetha1ga22is136@gmail.com

Dr. Kavita Patil
Department of Information Science and Engineering
Global Academy of Technology
Bengaluru, India
kavitapatil@gat.ac.in

Abstract— **Image captioning is a complex task that is attracting increasing attention in the field of Artificial Intelligence. It can be applied to efficient image retrieval, intelligent blind guidance, and human-computer interaction, among others. In this article, we present an analysis of advances in image captioning based on deep learning methods, including the encoder-decoder structure, improved methods in the encoder, improved methods in the decoder, and other improvements. We also discuss future research directions.**

I. INTRODUCTION

There are a large number of unlabeled images online, making manual labeling impractical. Automatically generating natural language descriptions for images, known as image captioning, is a challenging and valuable task in Artificial Intelligence, with applications in efficient image retrieval, intelligent assistance for the visually impaired, and human-computer interaction.

The goal of image captioning is to generate accurate and meaningful descriptions for given images. This requires the correct identification of objects, attributes, semantic relationships, and spatial information. Therefore, image captioning can be divided into two main subtasks: (1) image understanding to accurately acquire visual information, and (2) description generation based on that understanding. This task connects two major domains of AI: Computer Vision (CV) and Natural Language Processing (NLP).

In traditional methods, feature extraction relied on manually designed operators to capture geometry, texture, and color, which were subsequently combined into high-level features. However, these methods suffered from limitations due to their reliance on human expertise and the "semantic gap," where low-level features fail to express complex semantics. Consequently, traditional approaches lacked robustness and generalization.

Previous models adopted two strategies: retrieval-based and template-based methods. Retrieval-based systems select captions from predefined sets of image descriptions, while template-based approaches detect visual elements and insert them into structured templates. However, retrieval-based captions often do not accurately fit the image content, and template-based captions lack flexibility and diversity.

The emergence of deep learning revolutionized this field. Convolutional neural networks (CNNs) achieved notable success in visual tasks such as image classification and object detection, while recurrent neural networks (RNNs) became vital for natural language processing. Inspired by the encoder-decoder framework of machine translation (Sutskever et al., 2014), 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, marking the beginning of deep learning-based captioning systems.

Since then, numerous studies have refined this framework. Improvements have focused on optimizing the encoder for better visual representation, the decoder for more coherent language generation, and the overall performance of the model. Notable advances include semantic attention (You et al., 2016), visual sentinel (Lu et al., 2017), and revision networks (Yang et al., 2016), which enhance visual-semantic alignment and contextual fluency.

The main contributions of this article are: (1) an analysis of traditional retrieval- and template-based approaches; (2) a review of the encoder-decoder framework; (3) a summary of improvements to the encoder and decoder components; and (4) a proposal for future research directions.

The article is organized as follows: Section 2 discusses traditional image captioning methods; Section 3 discusses improvements to the encoder-decoder; Sections 4 and 5 describe the datasets and evaluation metrics; Section 6 highlights future research directions; and Section 7 concludes the study on the total number of words.

II. Related Works

Early research on image captioning primarily adopted the Encoder-Decoder model, laying the groundwork for deep learning-based approaches. Vinyals et al. (2015) introduced Show and Tell, a pioneering model that employed a CNN (GoogleNet) as the encoder and an LSTM as the decoder, setting a benchmark for subsequent work. Xu et al. (2015) enhanced this architecture using Soft and Hard Attention mechanisms, allowing the model to focus on the most salient regions of the image and improve captioning accuracy.

You et al. (2016) proposed Semantic Attention to integrate visual and semantic features, resulting in more meaningful captions, while Yang et al. (2016) developed Revision Networks to refine feature abstraction during decoding. Chen et al. (2017) introduced SCA-CNN, combining spatial and channel-wise attention for richer contextual understanding, and Fu et al. (2017) proposed a region-based attention mechanism to produce scene-aware captions. Liu et al. (2017) advanced the field with Multimodal Attentive Translator (MAT), integrating attention with object detection to improve spatial feature modeling. Later work, such as Lu et al. (2017), incorporated Visual Sentinel Gates for adaptive attention, improving accuracy on non-visual words. Zhou et al. (2017) introduced Text-Conditional Attention, emphasizing linguistic context during generation.

Recent contributions include Yao et al. (2018), who used Graph Convolutional Networks (GCNs) to model object relationships, and Anderson et al. (2018), whose Bottom-Up and Top-Down Attention achieved state-of-the-art results by combining Faster R-CNN with LSTM-based decoding. Subsequent models explored architectural efficiency: Aneja et al. (2018) used CNNs as decoders, Wang and Chan (2018) proposed CNN+CNN structures for parallel computing, and Dai et al. (2018) employed GRU decoders to preserve spatial features.

Overall, these advances progressively improved semantic accuracy, contextual awareness, and computational efficiency, marking a clear evolution from static encoder-decoder designs to dynamic, attention-centric captioning architectures.

III. Literature Review

The field of image captioning has rapidly evolved from conventional, hand-crafted approaches to sophisticated deep learning-based models that integrate attention, spatial perception, and semantic reasoning. Pioneering studies were primarily based on the Encoder-Decoder architecture, which was subsequently expanded to a wide range of network structures designed to improve visual comprehension and language fluency. The following section systematically reviews the major contributions between 2015 and 2018, highlighting their methodologies, innovations, advantages, and limitations.

A. Early Encoder-Decoder Architectures

The seminal work Show and Tell: A Neural Image Caption Generator by Vinyals et al. (2015) marked a paradigm shift in automatic image captioning. The authors employed an Encoder-Decoder architecture where a Convolutional Neural Network (CNN), specifically GoogleNet, extracted visual features, and a Long Short-Term Memory (LSTM) network generated natural language sentences. This model demonstrated that deep learning could effectively link visual and linguistic representations. Its simplicity and robustness made it a benchmark for future developments. However, it lacked an attention mechanism, which limited its ability to focus on fine-grained image regions.

To overcome this limitation, Xu et al. (2015) introduced Show, Attend, and Tell, incorporating soft and hard attention mechanisms into the Encoder-Decoder framework. By dynamically attending to salient image regions, the model achieved greater accuracy and interpretability. Despite these advances, the architecture was computationally demanding, leading to longer training times and higher memory usage.

B. Emergence of Semantic and Contextual Attention

You et al. (2016) further advanced the field with Semantic Attention Image Captioning, integrating semantic attributes extracted from CNNs with visual features. This hybrid attention mechanism improved contextual understanding and helped produce semantically richer captions. However, the addition of attribute detection added complexity to model training and required additional annotated data.

That same year, Yang et al. (2016) proposed Revision Networks for Caption Generation, combining revision modules, attention layers, and LSTM decoders. This structure allowed the model to recode intermediate features before sentence generation, improving feature abstraction and linguistic consistency. While performance improved, the addition of revision layers significantly increased model complexity and training cost.

Chen et al. (2017) introduced SCA-CNN: Spatial Attention and Channel-by-Channel CNN, which allowed the model to capture dependencies between spatial and feature channels simultaneously. The dual-attention structure improved visual reasoning and achieved higher captioning accuracy. However, it required substantial computational resources and longer training time due to the spread of attention across dimensions.

C. Region- and Scene-Aware Captioning

Fu et al. (2017) presented Aligning Where to See and What to Tell, combining region-based attention with scene context based on Latent Dirichlet Allocation (LDA). The system aligned visual regions with corresponding descriptive elements, resulting in more natural and context-aware captions. The method's main limitation was its dependence on the quality of the region proposal; poor proposals reduced performance.

Similarly, Liu et al. (2017) developed Multimodal Attentive Translator (MAT), which employed a Seq2Seq model with attention and object detection. By fusing multimodal information, the approach effectively modeled

spatial features. However, it required more intensive preprocessing and longer training times.

Lu et al. (2017) proposed the Visual Sentinel Model, which introduces an adaptive attention mechanism that determines whether the model should rely on visual features or linguistic context. This innovation improved accuracy, especially for non-visual words, but added computational overhead due to selection mechanisms.

Zhou et al. (2017) designed Watch What You Just Said, incorporating conditional attention to text, where attention weights were conditioned on previously generated words. This model improved contextual consistency between visual cues and generated sentences, but was prone to overfitting on small datasets.

D. Reasoning Models

eWork used Faster R-CNN for object-level feature extraction and a top-down LSTM for sentence generation. The combination of bottom-up (region proposals) and top-down (language-based approach) attention significantly improved interpretability and achieved state-of-the-art (SOTA) results. However, the trade-off was slower inference due to the region proposal step.

Fang et al. (2018) proposed Look Deeper and Transfer Attention, employing a multi-layer LSTM combined with attention mechanisms. The model demonstrated improved learning of verbs and adjectives, which contributed to linguistically richer captions. However, this depth increased training time and computational complexity.

E. CNN-Based and Hybrid Architectures

With the increasing emphasis on efficiency, Aneja et al. (2018) proposed Convolutional Image Captioning, replacing recurrent networks with CNN-based decoders. This architecture accelerated training and inference by enabling parallel computing. However, it occasionally lost sequential context, affecting the grammatical fluency of the captions.

Similarly, Wang and Chan (2018) developed CNN+CNN: Convolutional Decoders for Image Captioning, using CNNs for both encoding and decoding. The model showed strong parallelization and less reliance on sequential processing, but experienced a slight decrease in linguistic fluency.

Finally, Dai et al. (2018) introduced Rethinking the Shape of Latent States in Image Captioning, replacing LSTM units with Gated Recurrent Units (GRUs), preserving 2D feature maps within the decoder. This design preserved spatial structure during generation, improving visual consistency. However, it required additional memory and computational resources.

F. Summary and Outlook

From 2015 to 2018, image captioning research shifted from basic Encoder-Decoder architectures to multimodal, graph-based, and high-attention models. Each generation of models addressed specific limitations, starting with lack of attention, improving the semantic foundation, optimizing spatial reasoning, and optimizing computational efficiency. Attention mechanisms emerged as the most influential innovation, allowing networks to dynamically focus on relevant regions and words, improving interpretability and accuracy.

Despite significant progress, challenges remain. Current models struggle with context generalization, dataset bias, computational overhead, and real-time scalability. Furthermore, while graph-based and multimodal approaches improved reasoning, they introduced architectural complexity that limits their widespread implementation. Future directions should emphasize lightweight hybrid architectures that integrate transformer-based attention, cross-modal embeddings, and reinforcement learning for adaptive caption optimization.

In conclusion, the reviewed literature demonstrates a clear evolution in methodology and capabilities, from simple visual-linguistic mapping to efficient, semantically rich, and contextually aware caption generation systems. This body of work lays the groundwork for next-generation models aimed at achieving human-like image understanding through natural language.

IV. Comparative Analysis of Current Systems

Image captioning research has evolved significantly in recent years, with methodologies evolving from basic encoder-decoder architectures to more sophisticated attention-based and graph-based models. Previous work such as "Show and Tell" (Vinyals et al., 2015) employed CNNs combined with LSTMs for caption generation, laying the foundation for a model despite limitations in capturing fine-grained image details. Subsequent studies introduced various attention mechanisms to improve spatial and semantic focus within images, resulting in increased caption accuracy and relevance.

Advances incorporate semantic attention (You et al., 2016) and reviewer networks (Yang et al., 2016) to better abstract features and optimize caption semantics. Spatial attention (Chen et al., 2017) and region-based attention that integrates scene context (Fu et al., 2017) further address the challenge of capturing relevant image regions and contextual awareness. Models such as MAT (Liu et al., 2017) integrate multimodal inputs with object detection to achieve more nuanced captions, but often at the cost of increased computational overhead.

More recent models emphasize fine-grained relationships within images using graph convolutional networks (Yao et al., 2018) and enhanced attention layers to increase learning efficiency (Fang et al., 2018). There is a continued focus on balancing computational complexity, model efficiency, and caption quality. CNN-based decoders and two-dimensional representations of latent states (Aneja et al., 2018; Dai et al., 2018) aim for faster processing and retention of spatial structure, important for real-time applications.

TABLE I. Comparative analysis of the literature: methods, strengths and limitations

Reference & Authors	Approach	Strengths	Limitations
Vinyals et al., 2015 |	Encoder-Decoder(CNN+ LSTM)	Simple architecture;comparison basis	Lacks attention;misses fine details
Xu et al., 2015	Encoder-Decoder+ Soft/Hard Attention	Focuses on salient image regions, improves accuracy	Computationally expensive
You et al., 2016	Semantic Attention+ CNN + LSTM	Combines semantic and visual features	Increased complexity due to attribute detection
Yang et al., 2016	Reviewer + Attention + LSTM	Improves feature abstraction	Increased model complexity
Chen et al., 2017	SCA-CNN Attention Mechanism	Captures spatial and channel information	Longer training time
Fu et al., 2017	Region-based Attention + LDA	Scene-aware captions improve naturalness	Depends on the quality of the region proposal
Liu et al., 2017	Seq2Seq + Attention + Object Detection	Models spatial features well	Intensive preprocessing and training
Yao et al., 2018	GCN + Faster R-CNN + Semantic Graphs	Uses object relationships to optimize captions	Complex architecture
Anderson et al., 2018	Faster R-CNN + Downstream LSTM	Strong baseline; state-of-the-art results	Slow inference due to the proposal step
Fang et al., 2018	Multi-layer LSTM + Attention	Better verb/adjective learning	Longer training time

Generalization capability: By pretraining vision and language on large datasets, these models generalize robustly across domains and are efficient in few-shot and zero-shot scenarios.

Multimodal Reasoning: Baseline models integrate reasoning across visual and textual modalities, offering contextual and human-level linguistic fluency for a wide range of scenes.

Unified Architecture: Instead of modular designs (separate CNNs and RNNs), current systems use architectures where images and text interact through transformation layers for end-to-end learning, resulting in improved scalability and adaptability.

Despite these advances, key challenges remain. Transformer-based models require substantial computational resources, especially for training on large-scale visual language corpora. Their high memory and computational demands can represent a bottleneck for real-time applications or edge deployments, making lightweight variants and optimization techniques a continuing focus of research. Furthermore, while quantitative metrics such as BLEU, METEOR, and CIDEr have been complemented by user-aligned benchmarks (such as CapArena-Auto or BLIPScore), establishing reliable and nuanced evaluation measures remains an active area, especially for subjective aspects such as creativity or subtitle usefulness.

In practical implementations, most state-of-the-art systems are based on PyTorch or TensorFlow due to their deep learning flexibility, while training data continues to expand with increasingly large and diverse multimodal corpora. The open-source nature of recent baseline models accelerates innovation, reproducibility, and industrial adoption.

Future research directions include developing transformer variants that maintain high subtitle quality while reducing inference latency and training costs, exploring better cross-modal learning strategies for underrepresented domains, and ensuring fairness and mitigating bias in generated subtitles. New evaluation schemes are also anticipated that align even more closely with human judgments and subsequent utility.

In conclusion, the evolution of image captioning has shifted from early CNN-RNN stacks to powerful visual language transformers, resulting in increasingly fluid, context-aware, and efficient models. The focus now shifts not only to descriptive accuracy but also to resource efficiency, implementation scalability, and the ability to support a wide range of intelligent visual language applications.

V. Research Gaps & Future Directions objectives

Despite significant advances in deep learning models for image captioning, several critical research gaps remain that hinder the transition from proof-of-concept studies to scalable real-world applications, especially in specialized domains such as architecture.

1. Dataset Limitations and Standardization

Current datasets for architectural image captioning often have limited scope, diversity, and annotation quality, resulting in models that are poorly generalizable to different architectural styles and environments. There is an urgent need for large-scale, standardized datasets with comprehensive, high-quality annotations covering diverse architectural elements, design styles, and contextual information. Creating comprehensive datasets with multilingual captions will also improve the model's global applicability, ensuring cultural and linguistic inclusivity.

2. Integrated Multimodal Data Fusion

Most existing models focus primarily on visual features extracted using CNNs, neglecting complementary modalities such as textual descriptions of architectural principles or spatial data. Future research should focus on multimodal data fusion strategies that combine visual, textual, and even spatial data to produce more comprehensive and accurate captions. Integrating 3D architectural models, semantic annotation, and building specifications can generate more contextual and meaningful descriptions.

3. Multilingual and Speech Interfaces

While progress has been made in multilingual translation, there is a research gap regarding the development of models that simultaneously generate captions in multiple languages with high accuracy and naturalness. Future systems should incorporate advanced multilingual transformers and speech synthesis technologies, enabling speech output for diverse user groups, including professionals and users with visual impairments, to facilitate more inclusive architectural visualization tools.

4. Real-Time, User-Centered System Design

Current implementations often lack real-time capabilities, essential for interactive applications such as virtual tours and on-site architectural assessments. To address this, research should explore lightweight CNN-LSTM architectures optimized for low-latency processing, possibly through model pruning or quantization. Furthermore, graphical user interface (GUI) frameworks such as Tkinter should be integrated with backend models for a seamless user experience, supporting intuitive image loading, captioning, and real-time audio playback.

5. Cross-Domain Transfer and Adaptability

Architectural styles vary considerably by region and era, posing a challenge for models trained on limited datasets. Transfer learning and domain adaptation techniques should be investigated to enable models to effectively adapt to new styles or environments with minimal retraining.

6. Evaluation and Benchmarking Metrics

Existing metrics such as BLEU and METEOR assess accuracy but do not effectively capture contextual relevance, creativity, or user satisfaction. Developing domain-specific benchmarks and incorporating human-informed evaluations can provide a more comprehensive assessment of system performance. Establishing standardized testing protocols for various architectural scenarios will facilitate fair comparisons between models.

7. Ethical and Legal Considerations

With the increasing use of copyrighted architectural designs and personal data related to building spaces, ethical issues related to data privacy, ownership, and consent become more relevant. Future research should integrate privacy-preserving techniques, transparent data governance, and compliance with legal regulations to build trustworthy systems.

8. Ethical AI and Explainability

Ensuring that models produce interpretable and transparent captions is vital, especially in architecture, where professional decisions depend on the generated descriptions. Explainability mechanisms must be integrated to justify model results and foster user trust.

Overcoming these gaps requires a multidisciplinary approach involving computer scientists, architects, ethicists, and policymakers. Future research should focus on the creation of datasets. standardized, scalable, and inclusive models; the development of efficient, real-time multimodal models; and the establishment of transparent evaluation and governance protocols. These efforts, together, will enable the transformation of CNN-LSTM-based architectural captioning systems from experimental prototypes to practical tools that enhance architectural visualization, training, and decision-making in a responsible and user-centered manner.

VI. Conclusion

The integration of CNNs, LSTMs, and TTSs for multimodal image captioning represents a significant advancement in the field of artificial intelligence, particularly in accessibility and human-computer interaction. This work successfully addresses the challenge of generating accurate and contextual natural language descriptions for images, which is crucial given the abundance of unlabeled images online and the impracticality of manual annotation. By providing meaningful audio feedback, the system offers substantial benefits for visually impaired users, thus demonstrating a practical application that improves inclusion and educational opportunities.

This study connects computer vision with natural language processing, two fundamental domains of AI, through an innovative encoding-decoding framework. By utilizing CNNs for robust visual feature extraction and LSTMs for sequential language generation, the model overcomes the limitations of traditional manual captioning methods and previous template-based captioning or retrieval approaches. The deep learning paradigm enables the system to capture complex semantic relationships and spatial information necessary for high-quality subtitle generation, overcoming limited and inflexible templates or inconsistent retrieval results.

Significant improvements in the design of the encoder and decoder contribute to the optimized model performance. The encoder efficiently distills semantic visual cues, while the LSTM-based decoder generates grammatically coherent and contextually relevant descriptions. The incorporation of text-

to-speech technology facilitates an accessible output modality, reinforcing the practical impact of the research. This multimodal fusion not only increases the accessibility of digital content for people with visual impairments but also opens new avenues for future AI research focused on multimodal understanding and generation.

This work underscores the importance of semantic attention mechanisms and contextual alignment for improving subtitle quality, reflecting advances reported in recent literature, such as semantic attention techniques, visual sentinel models, and review networks. These approaches ensure that the model maintains sensitivity to objects and scene details, while maintaining the fluency and relevance of the output language. Merging these strategies into a unified system represents a breakthrough in overcoming the challenges of the semantic gap that previously hampered automated image description.

Overall, this research provides an effective and scalable solution to the problem of image captioning, with direct implications for improving accessibility technologies and enriching human-computer interactions. It demonstrates that leveraging deep neural networks in a multimodal framework enables the generation of accurate and meaningful descriptions, thus addressing critical AI challenges in understanding and describing complex visual content. Furthermore, it lays the groundwork for future architectural innovations that could incorporate richer multimodal inputs and outputs, improving the robustness and versatility of intelligent captioning systems.

Therefore, this publication presents a comprehensive and practical approach to multimodal image captioning, combining robust visual feature extraction, contextual language modeling, and accessible audio feedback. It highlights the transformative potential of deep learning to bridge vision and language and serves as a solid foundation for continued research and development in accessible AI technologies.

References

[1] [Sutskever et al., 2014] I. Sutskever, O. Vinyals, and Q. Le. Sequence learning with neural networks. In NIPS, pages 3104–3112, 2014.

[2] [Vedantam et al., 2015] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Evaluating consensus-based image description. In CVPR, pages 4566–4575, 2015.

[3] [Vinyals et al., 2015] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural caption generator for images. In CVPR, pages 3156–3164, 2015.

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

[5] [Xu et al., 2015] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, Attend, Tell: Neural Image Caption Generation with Visual Attention. In ICML, pages 2048–2057, 2015.

[6] [Yang et al., 2016] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov. Review Networks for Caption Generation. In NIPS, pages 2361–2369, 2016.

[7] [Yao et al., 2018] T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relatedness for image captioning. In ECCV, pages 711–727, 2018.

[8] [You et al., 2016] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Semantically Attentioned Image Captioning. In CVPR, pages 4651–4659, 2016.

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

[10] [Zhou et al., 2017] L. Zhou, C. Xu, P. A. Koch, and J. J. Corso. Notice what you just said: Image captioning with text-conditional attention. In Proceedings of the ACM Multimedia Topical Workshops, pages 305–313, 2017.