Review

A Comparative Survey of CNN-LSTM Architectures for Image Captioning

Sehran Sajad Bhat *, Shafin Mehnaz , Shadab Ali Shekh , Tasbeeha F. , Lijimol K.

Posted Date: 15 December 2025

doi: 10.20944/preprints202512.1301.v1

Keywords: image captioning; convolutional neural networks (CNN); long short-term memory (LSTM); deep learning; computer vision; natural language processing; survey

 doi:10.20944/preprints202512.1301.v1

Review

A Comparative Survey of CNN-LSTM Architectures for Image Captioning

Sehran Sajad Bhat *, Shafin Mehnaz, Shadab Ali Shekh, Tasbeeha F. and Lijimol K.

Department of Computer Science and Engineering, HKBK College of Engineering, Bangalore, Karnataka, 560045, India
* Correspondence: 1hk22cs139@hkbk.edu.in

Abstract

Image captioning, the task of automatically generating textual descriptions for images, lies at the intersection of computer vision and natural language processing. Architectures combining Convolutional Neural Networks (CNNs) for visual feature extraction and Long Short-Term Memory (LSTM) networks for language generation have become a dominant paradigm. This survey provides a comprehensive overview of fifteen influential papers employing these CNN-LSTM frameworks, summarizing their core contributions, architectural variations (including attention mechanisms and encoder-decoder designs), training strategies, and performance on benchmark datasets. A detailed comparative analysis, presented in tabular format, evaluates these works by detailing their technical approaches, key contributions or advantages, and identified limitations. Based on this analysis, we identify key evolutionary trends in CNN-LSTM models, discuss prevailing challenges such as generating human-like and contextually rich captions, and highlight promising future research directions, including deeper reasoning, improved evaluation, and the integration of newer architectures.

Keywords: image captioning; convolutional neural networks (CNN); long short-term memory (LSTM); deep learning; computer vision; natural language processing; survey

1. Introduction

Image captioning, the task of generating a natural language description corresponding to the visual content of an image, represents a fundamental challenge bridging the fields of computer vision (CV) and natural language processing (NLP). Its success holds significant potential for applications ranging from aiding visually impaired individuals and enhancing image retrieval systems to facilitating human-computer interaction. Over the past decade, deep learning techniques have driven remarkable progress in this area.

A pivotal development was the adoption of encoder-decoder frameworks, typically employing Convolutional Neural Networks (CNNs) as powerful visual encoders to extract rich image features, and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, as decoders to generate sequential textual descriptions [1]. This CNN-LSTM paradigm rapidly became the cornerstone for many state-of-the-art image captioning models, demonstrating impressive results on benchmark datasets.

Given the proliferation of variations and improvements built upon this core architecture (such as attention mechanisms [2], adaptive decoding, and integration of semantic concepts), a comprehensive synthesis of these approaches is essential. This survey aims to provide a structured overview and critical comparison of the key advancements within the CNN-LSTM based image captioning landscape. We focus specifically on understanding the evolution, strengths, and limitations of these influential models.

The scope of this survey encompasses 15 significant papers primarily utilizing CNNs for image encoding and LSTMs for caption generation, published roughly between 2011 and 2024. While acknowledging the rise of Transformer-based models, this review concentrates on the foundational

 doi:10.20944/preprints202512.1301.v1

and refined CNN-LSTM architectures to provide a focused analysis. We examine variations including different CNN backbones, LSTM cell modifications, attention mechanisms, training objectives, and evaluation strategies. Papers focusing solely on video captioning or employing non-deep learning methods are excluded.

The main contributions of this survey are: (i) a categorization of CNN-LSTM based image captioning models based on architectural innovations and learning strategies (though presented sequentially here); (ii) a concise summary of the core ideas and contributions of each selected paper; (iii) a comparative analysis evaluating the models , computational considerations, and qualitative aspects of caption generation; and (iv) the identification of key research trends, persistent challenges, and promising future research directions within this specific domain.

The remainder of this paper is organized as follows: Section 2 provides brief background on core concepts and evaluation metrics. Section 3 presents the sequential review of the selected literature. Section 4 offers the comparative analysis tables. Section 5 discusses trends, challenges, and future directions derived from the analysis. Finally, Section 6 concludes the survey.

2. Background and Preliminaries

Image captioning aims to generate a textual description $S = (w_1, w_2, \ldots, w_N)$ for a given image I. The dominant deep learning approach, and the focus of this survey, employs an encoder-decoder architecture. Typically, a pre-trained Convolutional Neural Network (CNN) acts as the encoder, extracting a rich visual feature vector or spatial feature map from the image. A Long Short-Term Memory (LSTM) network then serves as the decoder, generating the caption word-by-word, conditioned on the visual features and the previously generated words. Many refinements incorporate attention mechanisms, allowing the LSTM decoder to dynamically weight the importance of different image regions during generation [2]. Inference commonly uses beam search decoding to improve caption quality over greedy approaches.

Comparing captioning models requires standard quantitative metrics. Widely adopted metrics include: **BLEU** [3], measuring n-gram precision against reference captions; **METEOR** [4], incorporating synonymy and stemming for better semantic matching; and **CIDEr** [5], which measures consensus using TF-IDF weighting of n-grams and is often well-correlated with human judgment. Other metrics like ROUGE [6] and SPICE [7] are also sometimes reported. Evaluation is typically performed on benchmark datasets such as Flickr8k [8], Flickr30k [9], and MS COCO [10].

3. Literature Review

This section summarizes the key aspects of the 15 selected papers.

Paper 1 [11]: Maheswari et al. address the problem of automatically generating descriptive captions for accessibility by implementing a standard image captioning model. Their approach combines a ResNet50 CNN for feature extraction with an LSTM network for sequence generation, utilizing GloVe embeddings. The work involves standard preprocessing and training steps and serves as a demonstration of this common deep learning paradigm for the image captioning task.

Paper 2 [12]: Kinghorn et al. tackle the issue that holistic captioning methods may overlook important local details, leading to less descriptive captions. They propose a region-based pipeline starting with an R-CNN object detector, followed by separate LSTMs for predicting human/object attributes and a CNN for scene classification. An encoder-decoder LSTM then translates these detected elements into refined, descriptive sentences. The authors claim their method generates more detailed captions by focusing on local regions and demonstrate outperformed contemporary baselines on the IAPR TC-12 dataset, also showing strong cross-domain performance on NYUv2.

Paper 3 [13]: Yuan et al. address the potential information loss when using only global or only local image features and potential timescale limitations of LSTMs. They propose the "3G" model, which fuses global (VGG FC7) and local (VGG Conv5-4 + attention) features via an adaptive gate. For sequence modeling, they utilize a 2-layer Gated Feedback LSTM (GF-LSTM). Their contribution lies in

 doi:10.20944/preprints202512.1301.v1

combining global/local information adaptively and introducing GF-LSTM to captioning for potentially better handling of dependencies, reporting strong benchmark results.

Paper 4 [14]: Sasibhooshan et al. focus on extracting finer-grained semantic details and contextual spatial relationships for richer captions. Their approach employs a Wavelet transform-based CNN (WCNN) encoder, a Visual Attention Prediction Network (VAPN), and a Contextual Spatial Relation Extractor (CSE) module with an LSTM decoder, trained using CIDEr optimization. Key contributions include the novel WCNN features, the combined attention mechanism, and the explicit modeling of spatial relations, leading to high reported CIDEr scores.

Paper 5 [15]: Verma et al. present work on the standard image captioning task, using a VGG16 Hybrid CNN (pre-trained on objects and scenes) as the encoder and a standard LSTM decoder without explicit attention. They highlight the use of the hybrid CNN and report competitive multi-metric results on standard benchmarks, including live image validation, though limitations in caption grammar/detail were noted.

Paper 6 [1]: Vinyals et al. present the foundational "Show and Tell" model, pioneering the end-to-end CNN-LSTM sequence-to-sequence approach. Using a GoogLeNet encoder to initialize an LSTM decoder and training via maximum likelihood, this work established the paradigm, showed significant BLEU improvements, and demonstrated the generative capabilities of these neural models.

Paper 7 [16]: Lu et al. explore the specific challenges of remote sensing (RS) image captioning. Their contribution is the creation and release of the large-scale RSICD dataset. They benchmarked standard captioning models on this dataset, identifying RS-specific difficulties and demonstrating the limitations of standard models in this domain.

Paper 8 [17]: Baig et al. address the description of novel objects not seen during training. They propose a modular post-processing method using an external object detector (YOLO9000) and Word2Vec embeddings to identify and then substitute nouns in a base-generated caption based on semantic similarity, improving scores on modified captions without retraining the captioner.

Paper 9 [18]: Zhang et al. utilize a Bidirectional LSTM (Bi-LSTM) decoder to incorporate future context and address potential state misalignment. Their approach uses a ResNet encoder with visual attention and introduces a "Subsidiary Attention" mechanism to fuse forward/backward Bi-LSTM states, reporting improved CIDEr performance on MS COCO.

Paper 10 [19]: Ming et al. provide a comprehensive review and taxonomy of the automatic image captioning field up to early 2022. Their work surveys traditional and deep learning methods, datasets, metrics, state-of-the-art comparisons, challenges, and future research directions.

Paper 11 [20]: Bai and An offer an earlier survey (up to 2018), classifying image captioning approaches into retrieval-based, template-based, and various neural network categories. They discuss strengths, limitations, benchmark results, and future directions, focusing primarily on neural methods.

Paper 12 [21]: Feng investigates knowledge-lean caption generation for news images using noisy web data. This early work uses LDA topic models on images and associated articles for content extraction, followed by extractive and abstractive (phrase-based) generation methods, demonstrating feasibility without manual resources.

Paper 13 [22]: Khademi and Schulte propose a hierarchical, context-aware architecture to improve captioning. They use BiGrid LSTMs for spatial context, integrate region-based text features, and employ a deep Bi-LSTM with dynamic spatial attention implemented via another Grid LSTM, marking the first use of Grid LSTMs for captioning and achieving strong results.

Paper 14 [23]: Arasi et al. focus on improving performance by optimizing hyperparameters using metaheuristics. Their proposed AIC-SSAIDL technique combines a MobileNetv2 encoder tuned with Sparrow Search Algorithm (SSA) and an Attention Mechanism-LSTM decoder tuned with Fruit Fly Optimization (FFO).

Paper 15 [24]: Amirian et al. present a concise review emphasizing the algorithmic overlap between deep learning methods for image and video captioning. They discuss shared architectures (CNNs, RNNs/LSTMs, GANs), datasets, metrics, and platforms relevant to both tasks.

 doi:10.20944/preprints202512.1301.v1

4. Comparative Analysis

Tables 1 and 2 present a detailed comparative analysis of the fifteen seminal and recent papers identified in our literature review. To facilitate a clear and structured comparison, the core aspects of each study are distilled within these tables. For each paper, they highlight its primary authors, the specific technical approach and architectural choices made , the main contributions or advantages reported by the authors, and any significant limitations or trade-offs that were acknowledged.

Table 1. Comparative Analysis of Surveyed Papers (Selected Aspects) - Part 1 of 2.

Authors	Technical Approach	Contributions / Advantages	Limitations
Maheswari et al. [11]	ResNet50 encoder; LSTM decoder; GloVe word embeddings.	Implemented CNN+LSTM captioning model.	Standard architecture; Performance comparison needed.
Kinghorn, Zhang, Shao [12]	Region-based: R-CNN object detection, LSTMs for attributes, CNN scene classifier, Encoder-Decoder LSTM (labels to sentence).	More detailed captions via local focus; Outperformed baselines (IAPR TC-12); Cross-domain success (NYUv2).	Lower ROUGE-L score; Struggles with complex scenes; Processing time.
Yuan, Li, Lu [13]	Fuses global (VGG FC7) & local (VGG Conv5-4 + attention) features; Adaptive global gate; 2-layer Gated Feedback LSTM.	Combines global/local adaptively; Gated Feedback LSTM enhances language model; Strong benchmark results.	Outperformed by methods using external detectors; Minor caption errors.
Sasibhooshan, Kumaraswamy, Sasidharan [14]	Wavelet CNN encoder; Visual Attention Prediction Network (atrous, channel+spatial attention); Contextual Spatial Relation Extractor; LSTM decoder. Train: Cross Entropy + Self-critical (CIDEr optimization).	Novel Wavelet CNN features; Combined channel/spatial attention; Explicit spatial relation modeling; High CIDEr score (MS COCO).	Fails with complex scenes or incorrect object/relation recognition.
Verma, A. Yadav, Kumar, D. Yadav [15]	Encoder-Decoder: VGG16 Hybrid Places 1365 encoder, standard LSTM decoder. No explicit attention reported.	Hybrid CNN (objects+scenes); Multi-metric results reported; Claims competitive performance; Live image validation.	Captions lack grammar/ detail; High training time; Failure cases noted; Basic architecture (Preprint).
Vinyals, Toshev, Bengio, Erhan [1]	Encoder-Decoder: CNN (GoogLeNet) features feed LSTM decoder initially. End-to-end training (max likelihood). Beam search inference.	Foundational sequence-to-sequence model (NIC); End-to-end trainable; Significant BLEU score gains; Demonstrated generative diversity.	Overfits small datasets; Many verbatim captions; Gap versus human evaluation.
Lu, Wang, Zheng, Li [16]	Benchmarked standard methods (RNN/LSTM, Attention-LSTM) on Remote Sensing (RS) image data.	Created/released RSICD dataset; Identified RS captioning challenges; Showed standard model limitations on RS data.	No new model proposed; Benchmarked models rated 'acceptable' on RS; Dataset duplications noted; Poor cross-dataset performance.
Baig, Shah, Wajahat, Zafar, Arif [17]	Post-processing: External object detector (YOLO9000) + Word2Vec identify novel objects; Replaced nouns in base caption using semantics.	Handles novel objects without retraining; Modular approach; Uses external detector/embeddings; Score improvement shown (modified captions only).	Post-processing, not end-to-end; Depends on detector accuracy; Evaluation focused on changed captions.

 doi:10.20944/preprints202512.1301.v1

Table 2. Comparative Analysis of Surveyed Papers (Selected Aspects) - Part 2 of 2

Authors	Technical Approach	Contributions / Advantages	Limitations
Zhang, Ma, Jiang, Lian [18]	CNN(ResNet) encoder + Visual Attention. Bidirectional LSTM decoder. Novelty: Subsidiary Attention fuses forward/backward states. Train: Cross Entropy + Self-critical (CIDEr optimization).	Bidirectional context generation; Novel state fusion mechanism (Subsidiary Attention); Improved performance (esp. CIDEr) versus standard Bi-LSTM/others (MS COCO).	Bidirectional LSTM increases parameters/latency; Potential high-frequency word bias.
Ming, Hu, Fan, Feng, Zhou, Yu [19]	Review Paper: Surveys Traditional (Retrieval, Template) & Deep Learning (Encoder-Decoder, Attention, Training) methods.	Comprehensive survey & taxonomy; Summarizes datasets/metrics; Compares state-of-the-art; Discusses challenges.	Not applicable (Review Paper).
Bai, An [20]	Review Paper: Classifies Retrieval, Template, Neural Network-based (Multimodal, Encoder-Decoder, Attention, Compositional, Novel Object) methods.	Survey/summary of image captioning (up to 2018); Compares state-of-the-art; Discusses future directions.	Not applicable (Review Paper).
Feng, Y. [21]	Knowledge-lean news captioning: LDA topic model (image+document) for keywords; Extractive & Abstractive (phrase-based) realization.	Uses noisy web data (BBC News); Joint visual-text topic model; Knowledge-lean generation demonstrated; Phrase-based abstractive method.	News domain focus; Relies on topic models; Predates deep learning era.
Khademi, Schulte [22]	Bidirectional Grid LSTM (spatial features) + Region Texts (transfer learning) -> Deep Bidirectional LSTM (Layer 1: context, L2: generation) + Dynamic Spatial Attention (Grid LSTM).	Novel context-aware architecture; Grid LSTM for spatial context/attention; Uses region texts; Hierarchical context/generation; State-of-the-art performance (MS COCO).	Model complexity significant; Needs pre-trained dense captioner.
Arasi et al. [23]	Encoder: MobileNetv2 + Sparrow Search Algorithm (hyperparameter optimization); Decoder: Attention Mechanism-LSTM + Fruit Fly Optimization (hyperparameter optimization).	Proposed AIC-SSAIDL technique; Uses metaheuristics (SSA, FFO) for hyperparameter tuning; Reported improved results.	Focus on metaheuristic optimization; Gains depend on optimization algorithm success; Computational cost unclear.
Amirian et al. [24]	Review (Image & Video): Concise review of Deep Learning methods (CNN, RNN/LSTM, GANs), focusing on algorithmic overlap.	Links image/video captioning methods; Discusses architectures, datasets, platforms; Included case study (video titles).	Concise scope, not comprehensive; Focus only on Deep Learning & overlap.

5. Discussion: Trends, Challenges, and Future Directions

The comparative analysis presented in Table 1 and Table 2 reveals notable trends, persistent challenges, and potential future directions in the development of CNN-LSTM based image captioning models.

5.1. Observed Trends

Examining the technical approaches across the surveyed papers (Tables 1 and 2), the CNN-LSTM encoder-decoder architecture serves as a consistent foundation, originating from seminal work like Vinyals et al. [1]. A clear trend is the diversification of the CNN encoder component, moving from earlier architectures like GoogLeNet [1] and VGG variants [13,15] towards ResNet [11,18] and more specialized networks such as Wavelet CNNs [14] or efficient backbones like MobileNetv2 [23]. Input feature handling also evolved, with trends towards fusing global and local features [13] or incorporating region-based information [12].

On the decoder side, while the standard LSTM remains prevalent [1,11,14,15,22,23], a significant trend involves exploring more complex recurrent units like Gated Feedback LSTMs [13], Bidirectional LSTMs [18,22], and specialized Grid LSTMs [22] to potentially capture richer temporal dependencies or context. Perhaps the most prominent trend visible in the tables is the increasing adoption and sophistication of attention mechanisms. Absent in early models [1], various forms appear later, including attention over local features [13], combined channel/spatial attention [14], dynamic spatial

 doi:10.20944/preprints202512.1301.v1

attention [22], and specialized attention for fusing Bi-LSTM states [18], highlighting its perceived importance for improving caption relevance. Trends towards integrating richer context (spatial, hierarchical, object-based) [12–14,18,22] and employing advanced training strategies (like RL-based optimization or metaheuristic hyperparameter tuning [14,18,23]) are also evident from the comparative data. Evaluation practices show a continued reliance on BLEU but an increasing trend towards reporting metrics like METEOR and CIDEr [1,14,16,18].

5.2. Open Challenges

The "Limitations" column in Tables 1 and 2 underscores several persistent challenges. Achieving consistent human-like caption quality remains elusive, with limitations cited regarding genericness [12], grammatical errors or lack of detail [15], and the known gap between automatic metrics and human perception [1]. The difficulty in accurately describing complex scenes with multiple objects or intricate relationships is another recurring theme [12,14].

Furthermore, the limitations highlight issues of data dependency and generalization. Overfitting [1], poor cross-dataset performance [16], reliance on dataset vocabulary limiting novel object description [17], and dataset quality issues [16] point to the need for more robust models or better data handling. Dependencies on external modules like object detectors [12,17] or pre-trained systems [22] also introduce potential points of failure or limit end-to-end applicability.

Finally, the trade-off between performance and efficiency is evident. Several approaches report increased model complexity, high training costs, or potential computational overhead as limitations associated with their advancements [13,15,18,22,23]. Balancing the drive for higher metric scores with practical deployability remains an open challenge. The limitations of evaluation metrics themselves [1] also continue to complicate the assessment of true progress.

5.3. Future Research Directions

Addressing the challenges revealed through the comparison points towards several important future research directions. Enhancing **scene understanding and reasoning** beyond the capabilities reflected in the surveyed works [12,14] is critical, potentially through integrating external knowledge sources, such as commonsense knowledge graphs, or explicitly incorporating structured representations like scene graphs during the captioning process. Developing architectures with improved spatial and causal reasoning capabilities is essential for generating captions that reflect a deeper understanding of the visual context, moving beyond surface-level object descriptions.

A second vital direction lies in developing **better evaluation metrics and methodologies**. The identified gap between automatic metrics like BLEU and human judgments of caption quality [1], along with the inability of current metrics to fully capture factual accuracy, semantic relevance, or creativity, necessitates further research. Future work should aim to create new automatic metrics, potentially learning-based, that demonstrate stronger correlation with human assessment across these multiple dimensions. Furthermore, exploring more scalable, cost-effective, and reliable protocols for human evaluation remains important for truly gauging progress in the field.

Third, improving model **generalization, robustness, and fairness** is paramount. Models need to move beyond the stylistic biases of specific benchmark datasets like MS COCO and perform well on diverse, real-world data, including specialized domains like remote sensing [16]. Research into domain adaptation, few-shot or zero-shot learning for image captioning could reduce the reliance on massive labeled datasets for every new domain. Addressing the challenge of describing novel objects or concepts more effectively than current post-processing approaches [17] is needed. Finally, continued research into identifying and mitigating societal biases (e.g., gender, race) learned from training data is an essential ethical consideration for developing responsible captioning systems.

6. Conclusion

This survey presented a comparative analysis, summarized in Tables 1 and 2, of fifteen key papers focused on image captioning using the prevalent Convolutional Neural Network (CNN) and Long

 doi:10.20944/preprints202512.1301.v1

Short-Term Memory (LSTM) encoder-decoder framework. The comparison highlighted the **different approaches** taken within this dominant paradigm, showcasing the evolution from foundational sequence-to-sequence models [1] to variants incorporating diverse CNN backbones, advanced LSTM structures (GF-LSTM, Bi-LSTM, Grid LSTM), sophisticated attention mechanisms, and richer context integration strategies.

The analysis also underscored **persistent limitations**, including challenges in generating high-quality, human-like captions, accurately describing complex scenes, the imperfections of standard evaluation metrics, difficulties with domain generalization and novel objects, and the recurring trade-off between model complexity and efficiency, as noted across various surveyed works [1,12,15,16,18,22].

Key **observed trends** included the increasing sophistication of both encoder and decoder components, the centrality of attention mechanisms in later models, and a growing focus on context and advanced training techniques. Looking forward, critical **future research directions** involve enhancing models' reasoning capabilities, developing more reliable evaluation methods, and improving generalization and robustness. While the CNN-LSTM framework has been instrumental, the path forward likely involves building upon these foundations while exploring insights and architectures from newer paradigms like Transformers and large pre-trained multimodal models to further advance the challenging task of automatic image description.

References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164.
2. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning (ICML); Bach, F.; Blei, D., Eds. PMLR, 2015, Vol. 37, *Proceedings of Machine Learning Research*, pp. 2048–2057.
3. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL). Association for Computational Linguistics, 2002, pp. 311–318.
4. Denkowski, M.; Lavie, A. METEOR Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT). Association for Computational Linguistics, 2014, pp. 376–380.
5. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.
6. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Association for Computational Linguistics, 2004, pp. 74–81.
7. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2016, pp. 382–398.
8. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. In Proceedings of the Journal of Artificial Intelligence Research, 2013, Vol. 47, pp. 853–899.
9. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. In Proceedings of the Transactions of the Association for Computational Linguistics, 2014, Vol. 2, pp. 67–78.
10. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2014, pp. 740–755.
11. Maheswari, A.; Kajal.; Selvameena, R.; Kumar, K.V.; Shekar, M.G.; Rahul, M.V. Image Caption Generator Using CNN and LSTM. *International Journal for Multidisciplinary Research (IJFMR)* **2024**, *6*. Accessed via IJFMR website/PDF.

 doi:10.20944/preprints202512.1301.v1

12. Kinghorn, A.; Zhang, L.; Shao, L. A Region-based Image Caption Generator with Refined Descriptions. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 450–454. https://doi.org/10.1109/ICIP.2017.8296322.
13. Yuan, Z.; Li, Y.; Lu, W. 3G Structure for Image Caption Generation. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019, pp. 6329–6334.
14. Sasibhooshan, R.; Kumaraswamy, R.; Sasidharan, S. Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction. *Multimedia Tools and Applications* **2023**, *82*, 28143–28167. https://doi.org/10.1007/s11042-023-14618-z.
15. Verma, V.; Yadav, A.; Kumar, A.; Yadav, D. Automatic Image Caption Generation Using Deep Learning. *arXiv preprint arXiv:2212.04531* **2022**.
16. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *56*, 132–145. https://doi.org/10.1109/TGRS.2017.2744871.
17. Baig, M.O.; Shah, S.Z.; Wajahat, I.; Zafar, A.; Arif, M. Image Caption Generator with Novel Object Injection. In Proceedings of the 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2). IEEE, 2021, pp. 1–5. https://doi.org/10.1109/ICoDT252791.2021.9439097.
18. Zhang, H.; Ma, L.; Jiang, T.; Lian, S. Image Caption Generation Using Contextual Information Fusion With Bi-LSTMs. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**, *33*, 1770–1782. https://doi.org/10.1109/TCSVT.2022.3223177.
19. Ming, Y.; Hu, N.; Fan, C.; Feng, F.; Zhou, J.; Yu, H. Visuals to Text: A Comprehensive Review on Automatic Image Captioning. *IEEE/CAA Journal of Automatica Sinica* **2022**, *9*, 1339–1365. https://doi.org/10.1109/JAS.2022.105734.
20. Bai, S.; An, S. A Survey on Automatic Image Caption Generation. *arXiv preprint arXiv:1804.04464* **2018**.
21. Feng, Y. Automatic Caption Generation for News Images. PhD thesis, School of Informatics, University of Edinburgh, Edinburgh, 2011.
22. Khademi, M.; Schulte, O. Image Caption Generation with Hierarchical Contextual Visual Spatial Attention. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 2056–2064.
23. Arasi, M.A.; Alshahrani, H.M.; Alruwais, N.; Motwakel, A.; Ahmed, N.A.; Mohamed, A. Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model. *IEEE Access* **2023**, *11*, 104633–104642. https://doi.org/10.1109/ACCESS.2023.3317276.
24. Amirian, S.; Rasheed, K.; Taha, T.R.; Arabnia, H.R. Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap. *IEEE Access* **2020**, *8*, 218386–218400. https://doi.org/10.1109/ACCESS.2020.3042484.