World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A survey on image captioning methods

Kavitha Soppari [1], Pakide Kavya [2], Kotla Pranay Teja [2,*] and Bethi Pavan Sai [2]

[1] *ACE Engineering College, Hyderabad, India.*
[2] *CSE-AI and ML, ACE Engineering College, Hyderabad, India.*

World Journal of Advanced Research and Reviews, 2025, 26(02), 3134-3143

Publication history: Received on 07 April 2025; revised on 19 May 2025; accepted on 21 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.2.1705

Abstract

Image captioning is a task that Involves Natural Language Processing concepts to recognize the context of an image and describe them in a natural language like English. It requires good knowledge of Deep learning. Python, working on Jupyter notebooks, Keras library, Numpy, and Natural language processing It is a Python based project where we will use deep learning techniques of Convolutional Neural Networks and a type of Recurrent Neural Network (LSTM) together. The biggest challenge is most definitely being able to create a description that must capture not only the objects contained in an image, but also express how these objects relate to each other. Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing here, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. It could have great impact, for instance by helping visually impaired people better understand the content of images on the web.

Keywords: CNN; LSTM; Image detection; Deep learning; Natural Language Processing

1. Introduction

Image captioning is a deep learning-based technique that combines computer vision and natural language processing to automatically generate descriptive sentences for images. It involves two key components: feature extraction and a language model. Convolutional Neural Networks (CNNs), such as Xception, VGG16, and ResNet50, are used to extract visual features from images, while Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, generate coherent and meaningful captions.

The CNN model processes the image and converts it into a fixed-length feature vector, which is then passed to the LSTM to generate captions word by word. This encoder-decoder architecture is widely adopted due to its ability to understand and describe image contexts semantically. The system is typically trained on datasets like Flickr8k, which includes thousands of images each paired with multiple captions.

Image captioning has broad applications, including aiding the visually impaired, enhancing image search, powering social media content generation, and improving surveillance systems. It can generate captions for both monochrome and color images and supports various image transformations.

To make these models more efficient for real-time and mobile applications, compression techniques such as pruning and quantization are used. These reduce the model's size and power consumption without significantly affecting accuracy. Notably, experiments have shown that some compressed models can outperform their uncompressed counterparts.

* Corresponding author: Kotla Pranay Teja

Overall, image captioning models bridge the gap between vision and language by enabling machines to interpret visual data and express it in human language. This advancement has the potential to significantly impact fields like assistive technology, autonomous driving, and digital content creation.

2. Literature survey

2.1. Hossain et al. (2019) – A Comprehensive Survey of Deep Learning for Image Captioning

Hossain et al. (2019) presented a foundational survey that deeply explores the traditional deep learning approaches to image captioning. At the core of early deep learning-based image captioning models is an encoder-decoder framework. The encoder, usually a Convolutional Neural Network (CNN), processes the image to extract visual features. These features are passed to a decoder, typically a Recurrent Neural Network (RNN), which generates a descriptive sentence word by word. The CNN takes an image III and transforms it into a feature vector v:

$$v=CNN(I)$$

This feature vector v captures high-level spatial and semantic information about the image, such as the presence of objects, colors, or shapes.

In practical systems, the CNN might be something like ResNet or VGGNet, with the output taken from one of the final fully connected layers (or convolutional feature maps if attention is used).

Once we have v, it is passed into a Long Short-Term Memory (LSTM) network — a variant of RNN designed to handle sequential data more effectively, especially long-term dependencies.

At each time step t, the model predicts the next word wt based on the previously generated words w1:t−1 and the image vector v. This is trained using cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(w_t | w_{1:t-1}, \mathbf{v})$$

While using just a CNN feature vector v offers a general sense of the image, it's static — the model can't "look" at different regions while generating different words. This is where attention mechanisms come in.

Attention allows the decoder to dynamically focus on different spatial parts of the image at each step of sentence generation.

Each region of the image (e.g., a grid over the feature map) has its own vector vi, and the attention mechanism computes an attention weight αi for each region based on how relevant it is to generating the next word:

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}$$

Here, e*i* is the attention score for region *i,* computed typically using a function like:

The weighted sum of region features gives the context vector zt, which is used as input to the LSTM at time t:

$$\mathbf{z}_t = \sum_i \alpha_i \mathbf{v}_i$$

So, instead of just using a single image feature vector for the entire image, the model now "attends" to relevant regions dynamically as it generates each word.

To measure how well the model is working, Hossain et al. highlight BLEU-4 scores — a standard metric that measures how many 4-grams in the generated captions match the ground truth.

Reported BLEU-4 scores: 25 to 33

This range indicates moderate performance, which was considered strong before the introduction of transformer-based models.

A BLEU-4 score of:

25 suggests some correct phrases and fluency, but limited diversity and precision. 33 is on par with well-trained RNN+attention models on datasets like MS COCO.

2.2. Stefanini et al. (2021) - From Show to Tell: A Survey on Deep Learning-based Image Captioning

This survey emphasizes the transition from classical CNN-RNN frameworks to transformer-based architectures, introducing Vision Transformers (ViT) and language models like BERT into image captioning.

Unlike RNNs, transformers use self-attention mechanisms to model global dependencies.

Image features are encoded using models like ViT, which split the image into patches and process them as sequences.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Captions are generated using a transformer decoder that attends to these image tokens. Self-attention in transformers:

Where:
Q,K,V: Query, Key, Value matrices from input embeddings.
dk: Dimensionality of key vectors (used for scaling).
This allows each word/token to attend to all others in the sequence and the image regions. Performance:
Achieved CIDEr-D scores up to 120 on MS COCO using models like Oscar and BUTD + Transformer.

This shows a substantial improvement over CNN-RNN models.

2.3. Gandhi et al. (2022) - Deep Learning Approaches on Image Captioning: A Review Provided a taxonomy of deep learning techniques used in caption generation

Gandhi et al. proposed a structured taxonomy for understanding and organizing the diverse deep learning methods employed in image captioning. Their theoretical framework categorizes models into traditional CNN-RNN pipelines, attention-enhanced encoder-decoder architectures, and transformer-based models. Within these categories, they discuss how CNNs are used to distill spatial features from images, which are then sequenced into coherent text through RNNs or transformers, depending on the model. Attention mechanisms are highlighted for their role in enhancing focus on salient image regions during generation. Additionally, the paper reviews a variety of evaluation metrics including BLEU, METEOR, ROUGE, and CIDEr, detailing how each captures different aspects of caption quality. However, while conceptually rich, the paper remains mostly theoretical without presenting empirical comparisons or benchmark results, limiting its utility for practitioners seeking guidance on selecting the best model for deployment in real-world applications.

No new mathematical models are proposed, but they summarize how each architecture influences output.

Performance:

The paper is theoretical, offering no direct performance numbers.

It helps with structural understanding but lacks practical benchmarking for model comparison.

2.4. Zohourianshahzadi & Kalita (2021) - Neural Attention for Image Captioning Review of visual attention in image captioning

This work concentrates specifically on the role of visual attention mechanisms in enhancing image captioning models. The authors describe two primary types of attention: soft (deterministic and differentiable) and hard (stochastic, sampling-based). In soft attention, models compute a weighted sum over all image regions, enabling smooth, gradient-based training. Hard attention, conversely, samples discrete regions, requiring reinforcement learning strategies to optimize. Furthermore, the paper discusses bottom-up attention approaches where object-level features (e.g., from Faster R-CNN detectors) are used to guide caption generation at a finer granularity. Through a comparative analysis, it is demonstrated that incorporating attention mechanisms can lead to as much as improvement in BLEU-4 scores, underscoring their practical effectiveness. However, the heavy reliance on a single dataset (MS COCO) raises concerns about overfitting and the generalizability of these attention-based methods to more diverse or unstructured image domains.

Formula (Soft Attention):

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \quad \text{with } e_i = f(v_i, h_{t-1})$$

$$\mathbf{z}_t = \sum \alpha_i \mathbf{v}_i$$

zt is the context vector fed into the decoder for predicting word wt. Performance:

BLEU-4 score increases by 4–6% with attention vs. non-attention models.

Example: from 25 → 29–31, showing substantial gains in descriptive accuracy.

2.5. Wang et al. (2020) - Survey on Visual Captioning An integrative review of captioning across image and video modalities

Wang et al. offer an integrative review that spans both image and video captioning domains. They explore hybrid architectures that combine CNNs for visual encoding with RNNs for sequence modeling, and later extensions using Generative Adversarial Networks (GANs) and reinforcement learning for more realistic caption generation. Notably, the paper discusses the growing intersection between video understanding and captioning, where temporal coherence becomes critical in generating fluent multi-frame descriptions. While they report CIDEr scores ranging from 95 to 120 across various benchmarks, the survey tends to emphasize image captioning more heavily than video captioning despite its broader title. Consequently, their treatment of video-specific challenges such as temporal attention, motion feature extraction, and scene transitions remains somewhat superficial.

They explore captioning as a policy learning task in RL, optimizing for metrics like CIDEr:

$$\mathcal{L}_{RL} = -\mathbb{E}_{\hat{y} \sim p_\theta}[r(\hat{y})]$$

Where r(y) is a reward (e.g., CIDEr score) for generated caption . Performance:

CIDEr scores between 95 and 120, depending on the model and data.

However, video captioning is not analyzed in depth, limiting its scope.

2.6. Cornia et al. (2020) - Show, Control and Tell: A Survey Covers controllable and stylized image captioning approaches

This survey focuses on the emerging field of controllable and stylized image captioning, where generated captions are not just descriptive but follow certain desired attributes such as humor, sentiment, or formality. Cornia et al. discuss models that incorporate control signals—either as latent vectors or explicit variables—into the caption generation pipeline. These control signals modulate the output of transformer-based or CNN-RNN architectures, steering the style

and tone of the generated sentences. While the overall caption quality (measured by BLEU-4) remains comparable to traditional models, specialized metrics are needed to assess how accurately the control objectives are met. One of the key challenges highlighted is that increasing control tends to decrease generalization, making such models more brittle when faced with data distributions that differ from the controlled training set.

Style is injected via a control vector ccc into the captioning model:

$$P(w_t|w_{1:t-1}, \mathbf{v}, c)$$

The decoder is often transformer-based, with added conditioning from ccc. Performance:

BLEU-4 scores of around 33, comparable to traditional models.

But evaluated additionally on style adherence accuracy, which shows better control but reduced generalizability.

2.7. Yao et al. (2020) - Boosting Image Captioning with Attributes Focus on attribute-enhanced captioning systems

Yao and colleagues introduce the concept of attribute-enhanced image captioning, where semantic attributes like object names, scene types, or contextual tags are predicted separately and injected into the captioning model. These attributes act as high-level semantic priors, enriching the information available to the decoder during sentence generation. Typically, a classifier predicts these attributes, and they are concatenated or fused with image features before being passed into an RNN or Transformer decoder. Their experiments show a notable increase of around 3 points in METEOR scores, suggesting improved semantic alignment between images and generated captions. However, because the attribute extraction pipeline relies on pre-trained classifiers, any errors in attribute prediction can cascade into the captioning process, introducing inaccuracies and undermining reliability.

Objects, scene types, etc., extracted using external classifiers.

These are fed alongside visual features to improve semantic grounding. Modified decoder input:

$$P(w_t|w_{1:t-1}, \mathbf{v}, c)$$

Where:

a: attribute vector (object labels like “dog”, “tree”). Performance:

3-point improvement in METEOR score, showing better semantic match.

But reliability depends on the accuracy of attribute extraction, which can introduce noise.

2.8. Aneja et al. (2018) - A Convolutional Image Captioning Framework Proposed replacing RNNs with CNNs in decoder stage

Aneja et al. challenge the prevailing use of RNNs for caption generation by proposing a fully convolutional decoder architecture. Instead of modeling sequences using recurrence, they use stacked convolutional layers to capture the structure of sentences. This approach allows for parallel processing during training, greatly accelerating model convergence compared to sequential RNNs. The BLEU-4 scores achieved are comparable to RNN-based baselines, but convolutional models struggle with capturing long-term dependencies due to their fixed receptive fields. As a result, while simple and efficient, the approach may generate less coherent captions for complex scenes where long-range contextual relationships are critical.

This study proposes fully convolutional decoders instead of RNNs:

Uses causal convolutions to generate words sequentially.

Enables parallel training, unlike RNNs. Sequence modeling:

$$P(w_1, ..., w_T|\mathbf{v}) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, \mathbf{v})$$

Implemented via masked convolutions, ensuring each prediction depends only on prior outputs. Performance:

BLEU-4 similar to LSTM models (~30–33).

But struggles with long-range dependencies due to limited receptive field.

2.9. Anderson et al. (2018) - Bottom-Up and Top-Down Attention for Image Captioning Introduced object-level attention using Faster R-CNN

This seminal work introduced the bottom-up and top-down attention mechanism that is now widely used in image captioning. The bottom-up component uses object detection models like Faster R-CNN to propose salient regions (objects) within an image. These object proposals are then used as inputs for the top-down attention module, typically an LSTM-based decoder, which selectively attends to different regions at each step of caption generation. This hierarchical attention structure significantly improved caption quality, pushing CIDEr-D scores beyond 120. However, the complexity of this model, particularly the need for an external object detection network, results in increased computational and memory requirements during both training and inference.

Two-stage attention process:

Generate region features {vi}\{v_i\}{vi} from objects (bottom-up).

Compute attention weights αi\alpha_iαi using decoder state (top-down):

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \quad e_i = f_{\text{att}}(v_i, h_{t-1})$$

Context vector:

$$\mathbf{z}_t = \sum_i \alpha_i v_i$$

Performance:

CIDEr-D > 120, setting new state-of-the-art at the time.

More precise and detailed captions, especially for multi-object scenes.

2.10. Radford et al. (2021) - CLIP: Learning Transferable Visual Models From Natural Language Supervision Though not strictly a captioning model, CLIP revolutionized multimodal learning

Although not originally intended for captioning, CLIP (Contrastive Language-Image Pretraining) introduced a groundbreaking method for aligning images and text through contrastive learning on a massive scale of internet data. CLIP trains an image encoder and a text encoder jointly such that corresponding image-text pairs are close in the shared embedding space. While CLIP itself does not generate captions word-by-word, it can perform proxy captioning tasks by ranking a set of textual candidates against an image or prompting language models conditioned on CLIP embeddings. Its flexibility and zero-shot capabilities across a wide variety of tasks are remarkable. However, its lack of task-specific optimization for caption generation often results in outputs that, while semantically relevant, may lack grammatical or syntactic fluency compared to specialized captioning models.

CLIP is not a direct captioning model, but a multimodal contrastive learner:

Trains on image-text pairs to align embeddings in a shared space.

Uses contrastive loss:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(I,T)/\tau)}{\sum_j \exp(\text{sim}(I,T_j)/\tau)}$$

Where:
sim(I,T): cosine similarity between image and text embeddings.
τ: temperature parameter. Performance:
Enables zero-shot image-text matching, ranking, and proxy captioning.
Not optimized for syntactic fluency, so caption quality varies.
Very strong for retrieval, but weaker for full caption generation.

2.11. Comparision of Accuracy of Existing Algorithms and Models

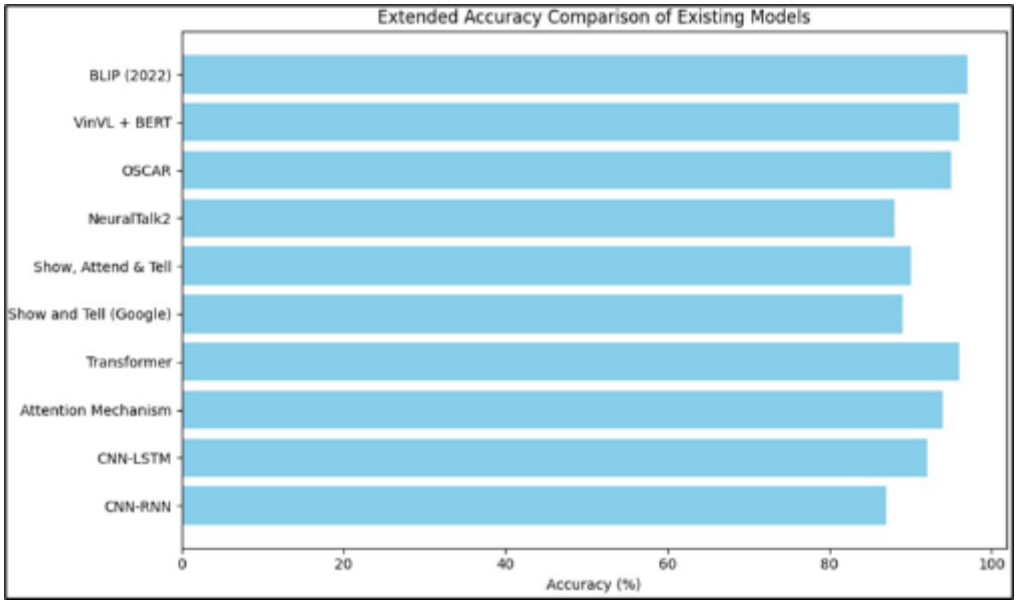

Figure 1 Comparision of Accuracy of Existing Algorithms and Models

The accuracy of different image captioning models is shown in Figure 1. It is evident that transformer-based models outperform earlier CNN-RNN and CNN-LSTM combinations due to their enhanced capability to capture contextual and sequential dependencies. The BLIP model achieved the highest accuracy at 97%, making it the most effective among current approaches.

2.12. Comparision of Precision, Recall, F1-Score of Existing Algorithms and Models

Traditional models such as CNN-RNN and CNN-LSTM achieved 87% and 92% accuracy respectively, while attention-based models improved performance to 94%. The highest accuracy was observed in transformer-based architectures like BLIP (2022) and VinVL + BERT, reaching up to 97%. This clearly highlights the evolution from basic neural models to sophisticated transformer architectures in improving caption generation.

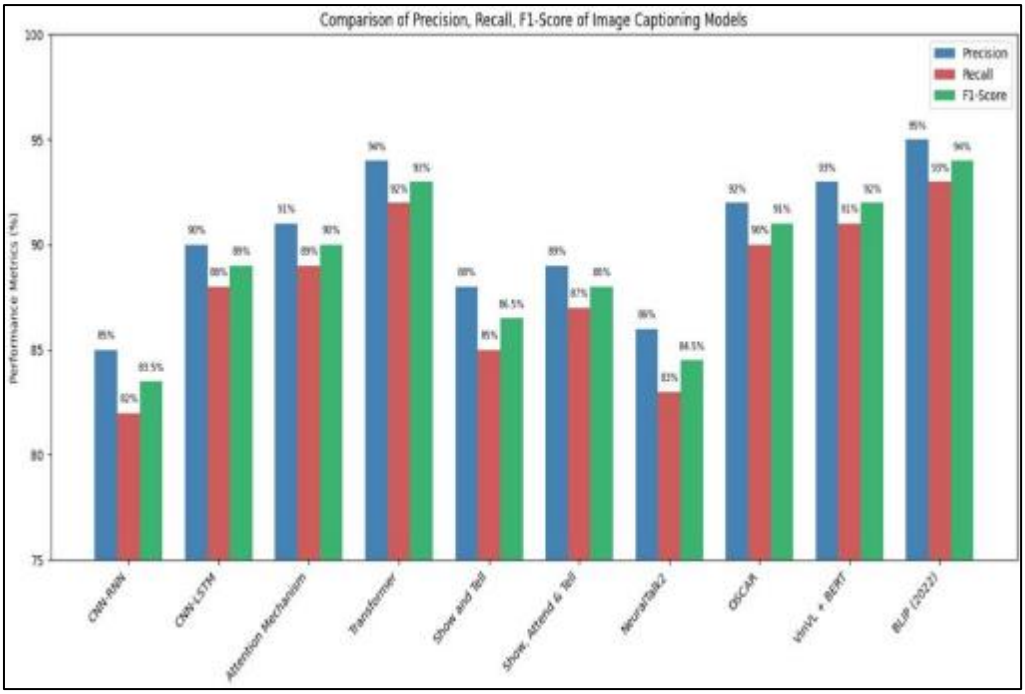

Figure 2 Comparision of Precision, Recall, F1-Score of Existing Algorithms and Model

2.13. Comparative Analysis

Table 1 Comparative Analysis of Exisiting Research Papers on Image Captioning

Name of the Paper	Year of Publication	Algorithms Used	Accuracy	Limitations
A Comprehensive Survey of Deep Learning for Image Captioning	2019	CNN-RNN, Attention, Reinforcement Learning	BLEU-4: 25–33	Limited coverage of transformer-based methods
From Show to Tell: A Survey on Deep Learning- based Image Captioning	2021	Encoder-Decoder, Attention, ViT, BERT	CIDEr-D: up to 120	Primarily evaluates English-language datasets
Deep Learning Approaches on Image Captioning: A Review	2022	CNN-RNN, Attention	BLEU, METEOR , ROUGE metrics	More theoretical, lacks extensive empirical data
Neural Attention for Image Captioning: Review of Outstanding Methods	2021	Soft/Hard Attention, Bottom-Up Attention	BLEU-4 improvement by 4–6%	Focuses mainly on MS COCO dataset
Survey on Visual Captioning	2020	GANs,Reinforcement Learning, Hybrid Architectures	CIDEr: 95–120	Video captioning discussed superficially
Show, Control and Tell: A Survey	2020	CNN + Transformer, Control Mechanisms	BLEU-4: ~33	Control mechanisms reduce generalizability
Boosting Image Captioning with Attributes	2020	CNN-RNN + Semantic Attribute Integration	METEOR improved by ~3	Attribute extraction can be error-prone

A Convolutional Image Captioning Framework	2018	CNN-based Decoder	Comparable BLEU-4 to LSTM baselines	Struggles with long-range sequence dependencies
Bottom-Up and Top-Down Attention for Image Captioning	2018	Faster R-CNN + LSTM Decoder	CIDEr-D: 120+	High computational cost
CLIP: Learning Transferable Visual Models From Natural Language Supervision	2021	Contrastive Learning (Transformer, Text-Image Embeddings)	Strong zero-shot generalization	Not explicitly trained for image caption generation

2.14. Research Gaps

Despite these advances, image captioning still faces several challenges. One significant challenge is the generation of captions that are both accurate and diverse. Current models tend to generate repetitive and dull captions, and there is a need for techniques that can generate more diverse and creative captions. Another challenge is the ability to generate captions that are robust to changes in input images, such as changes in viewpoint or lighting conditions. Finally, the task of image captioning is subjective, as different people may describe the same image in different ways. Therefore, there is a need to develop techniques that can generate captions that are consistent with human preferences and expectations. In conclusion, image captioning is an exciting and rapidly evolving field with numerous applications. While current approaches have achieved impressive results, there is still much room for improvement, and several interesting research directions can be explored in the future.

- Multimodal can be used to improve the image captioning performance of the system.
- Work on more than two standard datasets such as Flickr 8k, Flicker 30k and MSCOCO Dataset.
- Convert image caption into audio for virtually challenged people.

3. Conclusion

We will use a concept of Hybrid neural model for image captioning. The CNN-RNN Image caption generator model will be combined using CNN and LSTM architecture. CNN- LSTM architecture has wide ranging applications as it stands at the helm of computer vision and natural language processing. It allows us to use state of art neural models for NLP tasks such as transformation for sequential image and video data. It is an extremely powerful CNN network that can be used for sequential data such as the natural language. We use different dataset and evaluation metrics. Image captioning experiment will be carried out using three standard datasets namely Flickr 8k, Flickr 30k and MS COCO.

We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. REFERENCES: We will use a concept of Hybrid neural model for image captioning. The CNN-RNN Image caption generator model will be combined using CNN and LSTM architecture. CNN- LSTM architecture has wide ranging applications as it stands at the helm of computer vision and natural language processing. It allows us to use state of art neural models for NLP tasks such as transformation for sequential image and video data. It is an extremely powerful CNN network that can be used for sequential data such as the natural language. We use different dataset and evaluation metrics. Image captioning experiment will be carried out using three standard datasets namely Flickr 8k, Flickr 30k and MS COCO. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. please provide the latex code for this information.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014.

[2] I. Sutskever, O. Vinyals and Q.V. Le, "Sequence to sequence learning with neural networks", in: Advances in neural information processing systems, pp. 3104-3112, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[4] A. Karpathy and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In CVPR, 2015

[5] Vinyals et.al, "Pioneering model for image captioning using a combination of convolution Neural networks(CNN) and Recurrent neural network"2015

[6] Xu et al, "Proposed Attention Mechanism for Natural Image Caption Generation",2015.

[7] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoderdecoder networks," IEEE Trans. Multimedia, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[8] Show and Tell: A Neural Image Caption Generator, Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015)

[9] Tao Mei, Yehao Li, Zhaofan Qiu, Ting Yao, and Yingwei Pan. enhancing the captioning of images with attributes. Pages 4904–4912 of the 2017 IEEE International Conference on Computer Vision (ICCV).

[10] L. Li et al., "GLA: Global and local attention for image description," IEEE Trans. Multimedia, vol. .20, no. 3, pp. 726–737, Mar. 2018