A Comparative Study of Beam and Greedy Decoding Strategies for Image Captioning using Hybrid VIT-LSTM and Lightning Search Algorithm

Chandra Sekhar Sanaboina
Assistant Professor
Department of Computer Science and Engineering
University College of Engineering Kakinada
JNTUK - Kakinada – 533003
Andhra Pradesh

Girija Sankar Rotta
PG Scholar
Department of Computer Science and Engineering
University College of Engineering Kakinada
JNTUK – Kakinada – 533003
Andhra Pradesh

ABSTRACT

Image captioning, an interrelated task between computer vision and natural language processing, used to generate descriptive textual captions for given images. This paper presents an optimized deep learning-based Image Captioning System (ICS) that employs a Vision Transformer (ViT) as an image feature extractor and a Long Short-Term Memory (LSTM) neural network as the language decoder. To further enhance model performance, it incorporate the Lightning Search Algorithm (LSA), a nature-inspired metaheuristic algorithm, to automatically tune critical hyperparameters, including learning rate, dropout rate, and LSTM units. This automated optimization strategy improves both the quality of generated captions and the training performance. The proposed system is trained and evaluated on the Flickr30k dataset, achieving competitive performance across standard metrics such as BLEU, METEOR, and ROUGE. The results demonstrate that combining transformer-based vision encoders with recurrent language decoders, along with dynamic hyperparameter tuning algorithms, leads to more accurate and proficient image descriptions. This work contributes to the advancement of hybrid deep learning frameworks for image captioning tasks.

Keywords

Image captioning, Vision Transformer (ViT), Long Short Term Memory (LSTM), Lightning Search Algorithm (LSA), Deep learning, Hyperparameter optimization, Natural language processing, Beam Search, Greedy Search.

1. INTRODUCTION

Over the past few years, the exponential growth of digital images across various platforms, including social media, satellite systems, healthcare imaging, and real-time surveillance, has made it necessary to create intelligent systems that can comprehend and describe visual content. Image captioning—a significant task at the intersection of Computer Vision (CV) and Natural Language Processing (NLP)—automatically generates meaningful textual descriptions for images, thus facilitating applications in assistive technology, content-based retrieval, robotics, and remote sensing analysis [1], [2][3].

Earlier image captioning systems were built on handcrafted features or template-based rules that lacked flexibility, semantic richness, and the ability to scale across domains. With the advent of deep learning, encoder-decoder frameworks—particularly those combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks—have shown remarkable progress. These models could learn hierarchical visual features and generate coherent sentence sequences, achieving competitive scores on benchmark dataset such as Flickr30k [4], [5].

However, CNN-based encoders are inherently limited in modeling global spatial context and long-range dependencies, especially in complex scenes or aerial imagery. The drawback becomes important in such applications as remote sensing, where high-altitude imagery comprises macro-level and micro-level patterns in space. The best way to beat these challenges is the introduction of Vision Transformers (ViTs) that strongly compete with the self-attention capabilities that are used to capture relationships in the full image [6], [7]. Combined with sequence modeling models such as LSTM, ViT-LSTM bridges the gap between global attention and temporal modeling by having the advantage of ViTs and the benefit of LSTMs with hybrid ViT-LSTM models outperforming both models in context comprehension and linguistic expressiveness [8].

However, learning such models involves careful hyperparameter tuning (e.g. learning rate, dropout, size of hidden layer), which is currently often carried out by human experts, and is thus also quite likely to have sub-optimal hyperparameters. Such a bottleneck has prompted the application of meta-heuristic optimization such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO) and, of late, the Lightning Search Algorithm (LSA) [9]. Based upon the electrical branching characteristics of natural lightning, LSA is an effective method to sample high-dimensional parameter space and, therefore, converges more rapidly and with greater confidence in optimization of neural models [9],[10].

Simultaneously, many recent innovations have emerged in image captioning. Semantic attention mechanisms have improved image-text alignment by selectively focusing on context-relevant features[11], [12]. Scene graph-based methods extract relational structure among image objects to enhance caption diversity [13]. In OCR-rich datasets like TextCaps, multimodal Transformers using CLIP and visual-textual fusion have proven highly effective [14]. In remote sensing applications, dual-branch transformer encoders, graph convolution modules, and multi-level attention fusion are used to model complex spatial dependencies and attribute relations[15], [16], [17].

Moreover, reinforcement learning and non-auto regressive generation methods have been proposed to balance diversity, accuracy, and training efficiency [18]. POS-guided and linguistically informed captioning approaches offer further improvements in grammaticality and human-likeness of generated descriptions, especially for domains requiring structured language [19], [20].

Despite these advances, critical challenges persist—such as generating semantically grounded captions, achieving robustness to occlusion, and adapting to multi-domain generalization [21], [22], [23].

The proposed LSA-tuned ViT-LSTM image captioning framework addresses these challenges by combining the spatial precision of Transformers, the temporal fluency of LSTMs, and the optimization strength of LSA. Results from experiments on Flickr30k demonstrate that the model often surpasses baseline methodologies in metrics includes ROUGE, METEOR, and BLEU., while producing more informative, context-aware, and diverse captions[24].

2. RELATED WORK

Recent advancements in image captioning have been fueled by the synergy of deep neural architectures, attention mechanisms, semantic modeling, and optimization techniques. In order to contextualize ViT-LSTM image captioning system with the Lightning Search Algorithm (LSA) is proposed, The previous research papers are examined and they can be divided into four main categories: linguistic-semantic enhancements, transformer and attention-based architectures, optimization-based frameworks, and remote sensing-specific captioning.

2.1. Optimization-Based Captioning Frameworks

To overcome the manual trial-and-error method of hyperparameter tuning. Researchers are using metaheuristic optimization more and more to get around the manual and less-than-ideal nature of hyperparameter tuning in deep learning models. To improve generalization, one method combines a deep LSTM network with the Sparrow Search Algorithm (SSA) and Fruit Fly Optimization (FFO) to automatically modify important parameters like learning rate and dropout. The strategy put forth by Arasi et al. [10], demonstrates how well dual optimization works when negotiating intricate parameter spaces.

[25] introduced mg-BDRGRU, a bidirectional GRU decoder that uses depth residual connections for edge computing environments. The model shows how structural innovations can address both accuracy and latency, and is optimized for real-time inference on embedded hardware such as Jetson TX2.

A different approach that makes use of a multi-level deep reinforcement learning (RL) framework to optimize captioning policies at the word and sentence levels was put forth [18]. The model complements the LSA-based optimization by dynamically improving semantic consistency in captions by integrating vision-language and language-language rewards.

2.2. Transformer and Attention-Based Architectures

Transformers have reshaped sequence modeling, and their application to image captioning has proven highly impactful. An Adaptive Semantic-Enhanced Transformer, introduced in [6], employs weakly-supervised attention alongside adaptive gating to enrich semantic encoding. This design helps highlight meaningful regions in the image for more precise language generation.

To promote diversity and control in generated captions, [26] proposed a non-autoregressive, length-controllable transformer. Their model encodes target length as a guiding parameter and uses a refinement process akin to sequence-level knowledge distillation, thereby improving flexibility without compromising fluency.

Reward optimization has also been explored through hierarchical feedback mechanisms. [27] proposed a dual-network structure comprising a Revaluation Network (REN) and Scoring Network (SN) to evaluate sentence-level quality and correct word-level biases. Their approach aligns with the intention to generate more human-like, fluent captions.

[28] addressed the challenge of visual-semantic mismatch by designing a pyramid dual attention mechanism. Their method integrates spatial and channel-wise attention over multiple feature levels, enhancing the model's capability to distinguish fine-grained objects and their contextual relevance.

2.3. Remote Sensing Image Captioning (RSIC)

Remote sensing images are distinct in scale, texture, and content, requiring specialized captioning strategies. One such solution is Chg2Cap, a Siamese CNN-based captioning framework designed to detect and describe bi-temporal changes in satellite images. As described in [29], the system effectively employs attention to focus on altered regions between temporal image pairs.

The RSICCformer model, developed [17], which applies a dual-branch transformer architecture to remote sensing change captioning. Trained on the LEVIR-CC dataset, it separates content extraction and temporal fusion into parallel streams before merging the outputs through a dedicated decoder.

Based on the interpenetration of globals and locals, [8] came up with a framework of the neural network combining the benefits of Transformers and those of multi-scale feature extraction (MG-Transformer) that combines ResNet and CLIP embedding. Their solution utilizes Global Grouping Attention (GGA) and Meshed Cross-Attention (MCA) modules that provide the possibility of semantic alignment at the high-resolution level of satellite images.

[30] used the SA-FWC model, which works on small-scale data and with complex terrain using Sequential Attention and Flexible Word Correlation. It uses LSTM-based decoding together with self-attention-boosted features of VGG16 to maintain spatial leaders and consistent grammar inside the storytelling.

2.4. Semantic and Linguistic Enhancement Frameworks

There have been a few studies which have examined semantic alignment and linguistic guidance as caption quality requires seemingly more and more fluidity of human speech. [12] created a fine-grained attention mechanism that minimizes the odds of ambiguous or irrelevant captions through matching specific visual areas with semantic descriptors; it comes in handy in scenes having so many objects.

Due to the problem of vocabulary monotony, [19] proposed a Part-of-Speech (POS) Guidance Module. It enhances syntactic variety, but does not interfere with the semantic flow since it prefers words according to grammatical limitations.

[20] which employed multi-task learning to forecast the POS tags and produce captions simultaneously. In their way, they directly integrate linguistic structure into the decision-making

process of the decoder and make use of merge-based and inject-based methods.

Finally, there is the Re-Caption framework which is proposed in [31]. The model employs internal visual and semantic saliency maps to edit the original caption in order to enhance descriptive accuracy and point out larger regions that are aesthetically more important. This avoids the use of any external saliency tool.

3. PROPOSED MODEL

This proposed framework as depicted in Figure 1 combines three potent elements to produce meaningful image captions: a Vision Transformer (ViT) is used to encode images, an LSTM network is used to generate captions, and the Lightning Search Algorithm (LSA) is used to optimize important hyperparameters.

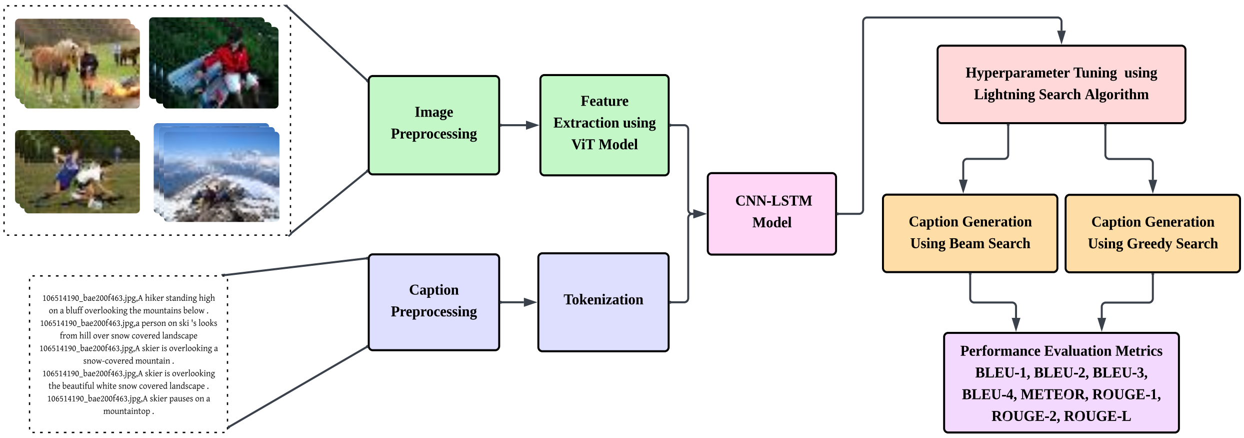
Image Preprocessing
Feature Extraction using ViT Model
Caption Preprocessing
Tokenization
CNN-LSTM Model
Hyperparameter Tuning using Lightning Search Algorithm
Caption Generation Using Beam Search
Caption Generation Using Greedy Search
Performance Evaluation Metrics BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L
106514190_bae200f463.jpg,A hiker standing high on a bluff overlooking the mountains below .
106514190_bae200f463.jpg,a person on ski 's looks from hill over snow covered landscape
106514190_bae200f463.jpg,A skier is overlooking a snow-covered mountain .
106514190_bae200f463.jpg,A skier is overlooking the beautiful white snow covered landscape .
106514190_bae200f463.jpg,A skier pauses on a mountaintop .

Figure 2. Workflow of Proposed System

These elements are tightly integrated into an end-to-end training system, and two decoding strategies are used to generate the final captions.

3.1. Image Encoder Using Vision Transformer (Vit)

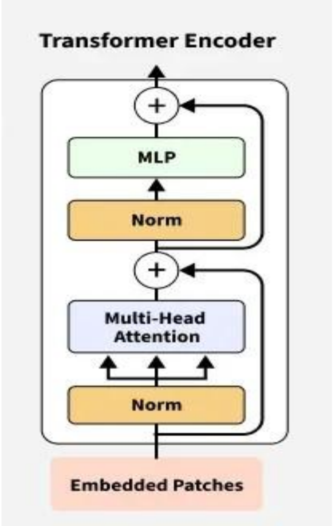

Figure 1. Vision Transformers Encoder [36]

The image encoder in this model as shown in Figure 2 is based on the Vision Transformer. It divides pictures into a number of fixed-size patches to process them. These patches are then flattened into vectors and projected into a higher-dimensional embedding space.

To preserve the spatial structure of the image, each patch embedding is enriched with positional encodings. After passing through several layers of transformer encoders, these vectors are subjected to multi-head self-attention mechanisms. This enables the model to capture global dependencies between various parts of the image as well as local features. The output is a comprehensive feature vector that represents the semantic content of the image. Compared to traditional CNN-based encoders, ViT offers greater flexibility and expressiveness, especially in complex or cluttered scenes.

3.2. Caption Generator Using LSTM

The Long Short-Term Memory (LSTM) network is used as the caption generator and is appropriate for sequential data modeling. It begins generating a caption one word at a time using the image features that the ViT encoder has produced.

At each step, the LSTM considers both the image features and the words that have already been generated. It maintains a hidden state that captures the context of the sentence so far, enabling it to produce fluent and grammatically correct outputs.

The predicted output at each step is passed through a dense layer, and the softmax activation function considers the most likely word from the vocabulary. The process repeats until an end-of-sequence token is predicted or the maximum caption length is reached.

3.3. Lightning Search Algorithm (LSA) for Hyperparameter Optimization

To enhance the performance and training efficiency of the model, the Lightning Search Algorithm is used to optimize key hyperparameters including,

- Learning rate
- Dropout Rate
- Batch Size

LSA begins by generating an initial population of random hyperparameter configurations, each referred to as a projectile. These candidates are evaluated by training the model briefly and measuring their validation loss. The configuration with the lowest loss is treated as the best-performing solution for that iteration.

In order to help the other candidates get closer to this ideal solution, updates are made later. The movement resembles the movement of energy in space, the movement of lightning in nature is recreated as well. To allow the algorithm go for a balance between exploration and convergence, the algorithm introduces a combination of adaptive steps, random perturbations and directional adjustments during updates.

Continuous hyperparameters are adapted on a per-step basis using Gaussian and exponential functions such as learning rate and dropout. Discrete parameters, like that of the batch size, however are adjusted slowly in order to achieve better values. On an LSA, one gets the LSA candidate configurations refined by several iterations until it picks the one that will bring the CG with the highest model accuracy and minimum validation loss. This automatic tuning process makes the model more resistant to changes and keeps the manual trial-and-error out of the equation.

3.4. End-to-End Training Workflow

The model is trained on a completely integrated worflow, and among them, there are ViT, LSTM, and LSA, which interact with each other to improve performance.

The ViT encoder extracts a high-level semantic feature representation then the ViT model takes the input image as inputs. After the input of the features, the LSTM decoder undertakes the autoregressive process of sequentially decoding captions based on the word-by-word prediction.

Categorical cross-entropy loss is used to compare the ground-truth captions and the predicted ones. Simultaneously the LSA evaluates the current configuration of hyperparameters and optimizes them once a better configuration is found. As the result of such dynamic adaptation, the decoder configuration and network weights are ensured to improve with time.

The quality of the produced captions is tested on the validation set with statistical measures e.g. BLEU, METEOR and ROUGE. With learning and optimization within a single loop, an effective model that can describe things that are syntactically fluent and semantically rich is generated in this workflow.

3.5. Caption Generation (DECODING) Strategies

The final captions are generated by the model after training through the process of decoding. The process of selecting these words in respect to the vocabulary at every level may considerably influence the quality of the descriptions generated. This paper uses Beam Search and Greedy Search to more the decoding algorithms.

3.5.1.Beam Search

Beam search as a decoder is more complicated, having several sets of candidate sequences at each time step. It expands every sequence of the top k (k being the beam width), accounting every possible candidate next-word instead of choosing a single word. The sequences are scored via cumulative log probabilities and only the highest scoring ones are retained [4].

This continues until a full sentence is arrived at. The diversity and fluency of the captions gets maximized significantly with the help of beam search since it explores a broader scope of possible choices. This model applies beam widths of three or five in effort to trade between cost and quality. Beam search is especially useful in images containing complex content and generating context aware and descriptive captions [19].

3.5.2.Greedy Search

Greedy search is a fast and easy decoding algorithm. At every time step, the model would pick the word in the softmax layer having the highest probability. This word is consumed as an input of the next time step until a maximum caption length is filled in or the end of sequence token is generated.

This is also a computationally efficient way to generate the captions that often turned out to be less imaginative or repetitive. The greedy search could lose the best overall sentence structure and get contented with the local optimality in a globally inferior outcome due to consideration of the most likely word at each step [26].

4. RESULTS AND DISCUSSION

A system whose components include the Intel Core i5 CPU, 16GB RAM, and 1TB of SSD is used to implement the proposed image captioning model, ViT-LSTM-LSA, based on Python 3.12.4. The Lightning Search Algorithm (LSA) rather than the manual tuning of training parameters dynamically generates key hyperparameters like batch size, learning rate, dropout rate etc.. LSA search can be thought of as an iterative method through which the hyperparameter area is searched and configurations updated using the validation loss to find the best settings at each training run. They set the training epochs to 50 and perform a nonlinear transformation in the model by using ReLU activation.

4.1. Dataset Details

The Flickr30k that is a well-established benchmark in the task of image to text generation and retrieval is chosen to evaluate the experimental results of the model ViT-LSTM-LSA. The FlickRelevance provides 31783 realistic photos that are in the Flickr collection. Each image comes with five different captions handwritten by native speakers of English and containing a diversity of detailed natural language explanations of the visuals. The elaborated details of the Flickr30k dataset is given in Table 1.

Table 1. Dataset Details

Dataset Name	Flickr30k
Total Images	31783
Captions per Image	5 (human-annotated)
Total Captions	1,58,915
Source	Flickr photo sharing platform
Language	English

According to every scene, the captions provide variety of linguistic expression which will allow models to acquire relationships of objects and action, contexts, and paraphrase. Consequently, the dataset is deemed to be particularly useful when it comes to evaluation of the capacity of an image captioning system to generate natural-sounding and generalized descriptions [1], [3].

Figure 3. Sample images in Flickr30k Dataset [13]

The factor that has contributed to its popularity among image captioning studies is moderate size dataset, and extensive linguistic variation, making Flickr30k both feasible to train and complex to model. The dataset is compatible with the modern deep learning frameworks and allows testing, validation, and training splits.

A sample of the images available in the Flickr30k dataset used in this research and containing multiple kinds of images depicting social relations, sporting activities, human actions and environment can be seen in Figure 3 .

4.2. Evaluation Metrics

On the Flickr30k dataset, the performance of the proposed captioning system which comprises the ViT (Vision Transformer), LSTM, and the Lightning Search Algorithm (LSA) to hyperparameter optimization was tested. The measurements are the quantitative ones: the BLEU-1 to BLEU-4 scores, METEOR, and ROUGE-1, ROUGE-2, ROUGE-L based comparison of results with the baseline architectures: ResNet, Inception, and Xception [19]. The evaluation item was based on an average of evaluation measures commonly adopted in natural language generation discipline:

4.2.1.BLEU (Bilingual Evaluation Understudy)

The n-gram precision between the generated and reference captions is evaluated using BLEU-1 to BLEU-4 scores. BLEU-1 measures unigram overlap, while BLEU-4 considers up to four-gram sequences, delivering a detailed analysis of syntactic accuracy and fluency. Higher scores indicate better alignment with human-generated captions.

4.2.2.METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR reviews captions based on synonym matching, stemming, word alignment, and semantic equivalence. This metric emphasizes both precision and recall and is effective in capturing linguistic variability and semantic relevance in the generated caption.

4.2.3.ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE-1 and ROUGE-2 calculate the overlap of bigrams and unigrams, respectively, and ROUGE-L calculates the LCS-Longest common sequence between the reference and candidate captions. The coherence of sentences and lexical similarity are reflected in these scores.

4.3. BLEU Score Analysis

Table 2. BLEU scores across various feature extraction Models

Model	**BLEU-1**		**BLEU-2**		**BLEU-3**		**BLEU-4**	
	Beam	**Greedy**	**Beam**	**Greedy**	**Beam**	**Greedy**	**Beam**	**Greedy**
ResNet	0.546	0.509	0.358	0.324	0.228	0.197	0.139	0.114
Inception	0.537	0.497	0.347	0.314	0.225	0.193	0.139	0.111
Xception	0.543	0.507	0.356	0.326	0.235	0.202	0.152	0.119
ViT	0.552	0.539	0.370	0.354	0.247	0.227	0.155	0.139

As displayed in Table 2, the ViT-based model achieved the highest BLEU scores across all n-gram levels using both greedy and beam search strategies. Specifically, BLEU-4 scores reached 0.155 (beam) and 0.139 (greedy) with ViT, outperforming Xception (0.152 / 0.119) and ResNet (0.139 / 0.114). The improvement is attributed to ViT's superior ability to model global visual dependencies, which enhances the quality of contextually appropriate word predictions.

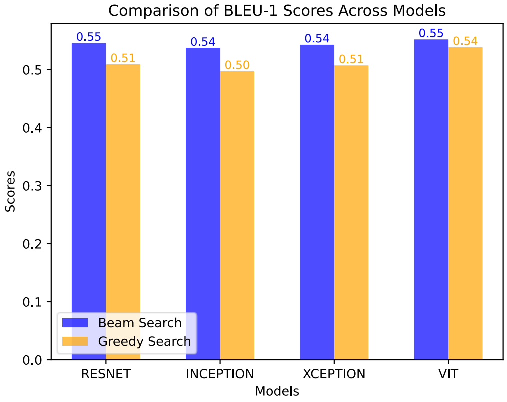

Figure 4. BLEU-1 Scores Comparison Across Models

On the Flickr30k dataset, BLEU-1 scores in figure 4 show that the ViT-based model with beam search achieves the highest unigram precision (0.552), slightly ahead of all other architectures, and even with greedy decoding it maintains a strong score (0.539). The small gap between beam and greedy results suggests that the approach generates accurate word choices inherently, without heavy reliance on search strategies. This highlights the strength of the ViT–LSTM–LSA framework in consistently selecting relevant words for captions, setting it apart as the most effective among the evaluated models for single-word accuracy.

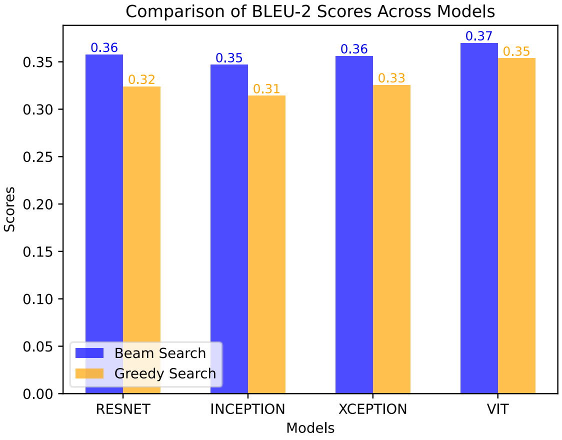

Figure 5. BLEU-2 Scores Comparison Across Models

In figure 5, BLEU-2 scores are plotted, which measures bigram precision and captures the accuracy of short word sequences, the ViT–LSTM–LSA model records the highest score with beam search (0.370), outperforming all other evaluated architectures. Even under greedy decoding, it achieves a strong score of 0.354, maintaining a clear lead over competing models. The relatively small gap between beam and greedy decoding indicates that the model's feature extraction and language generation components are inherently effective at producing coherent word pairs without heavy reliance on advanced search strategies. This superior performance in BLEU-2 demonstrates the model's capability to generate captions that not only contain accurate individual words but also form meaningful and contextually relevant short phrases.

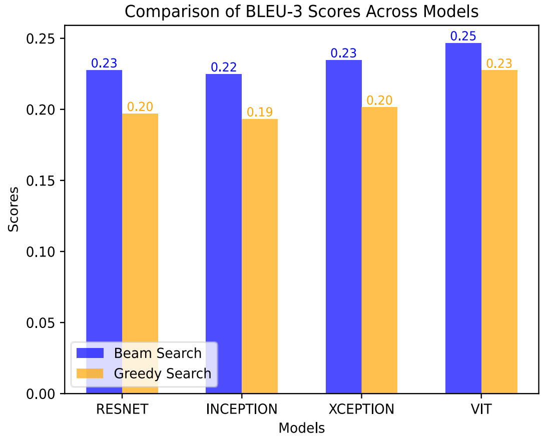

Figure 6. BLEU-3 Scores Comparison Across Models

In figure 6, it shows thef BLEU-3 scores, which measures the precision of three-word sequences, the ViT–LSTM–LSA model achieves the highest score with beam search at 0.247, followed by Xception beam at 0.235, ResNet beam at 0.228, and Inception beam at 0.225. Under greedy decoding, ViT records 0.227, maintaining its lead over Xception (0.202), ResNet (0.197), and Inception (0.193). The beam–greedy gap for ViT is 0.020, smaller than that of the CNN-based models, indicating consistent performance across decoding strategies. These results demonstrate the model's ability to retain accuracy in longer n-grams while delivering top performance in both search settings.

In figure 7, It shows the BLEU-4 scores, that it evaluates the precision of four-word sequences and is a stronger indicator of overall caption fluency, the ViT–LSTM–LSA model attains the highest score with beam search at 0.155, followed by Xception beam at 0.152, Inception beam at 0.139, and ResNet beam at 0.139. In greedy decoding, ViT reaches 0.139, ahead of Xception (0.119), ResNet (0.114), and Inception (0.111). The beam–greedy difference for ViT is 0.016, smaller than the gaps observed in the CNN-based baselines, reflecting its stability across decoding strategies. These results confirm the model's capability to maintain coherent and contextually accurate longer sequences, reinforcing its effectiveness in generating high-quality captions.

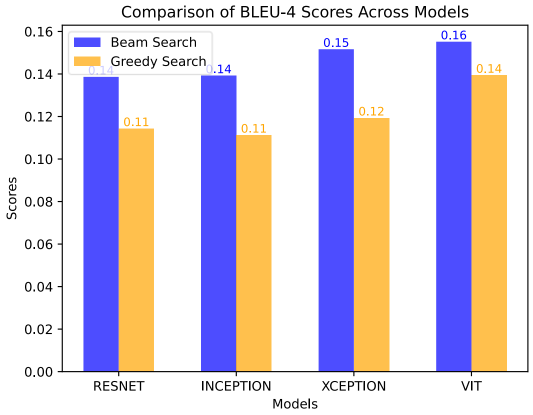

Figure 7. BLEU-4 Scores Comparison Across Models

The BLEU-n results (n = 1 to 4) show that the ViT–LSTM–LSA model performs better than all CNN-based models across every evaluation level, from single words to longer four-word sequences. It consistently achieves the highest scores with both beam search and greedy decoding, with only small differences between the two, showing that its predictions are accurate even without complex search strategies. The advantage becomes more noticeable in BLEU-3 and BLEU-4, where capturing longer and more meaningful phrases is crucial. Beam search improves results for all models, but the ViT-based approach remains the most effective overall, demonstrating a strong ability to generate captions that are both accurate and contextually relevant.

4.4. METEOR Score Comparison

Table 3. METEOR scores across various feature extraction Models

Model	Beam	Greedy
ResNet	0.360	0.338
Inception	0.362	0.334
Xception	0.356	0.324
ViT	0.376	0.356

Table 3 compares METEOR scores, which consider semantic matching and synonym handling. The ViT-LSTM-LSA model achieved 0.376 (beam) and 0.356 (greedy), surpassing the scores from all baseline models. These results validate that the ViT–LSTM–LSA system generates more meaningful and semantically aligned captions.

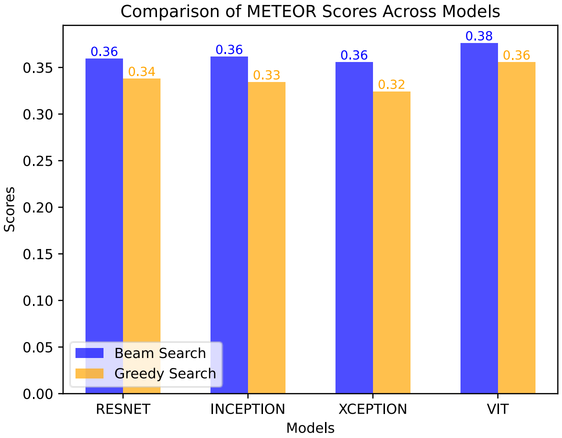

Figure 8. METEOR Scores Comparison Across Models

Figure 8 presents METEOR scores, which incorporate synonym matching, word stemming, and alignment-based penalties for more nuanced caption evaluation. The ViT-LSTM model achieved the highest METEOR score under both decoding strategies, reinforcing its advantage in generating semantically rich and linguistically precise captions. Compared to ResNet and Inception, the ViT model demonstrates better generalization in word choice and sentence formation, attributes critical for applications requiring natural human-like descriptions.

4.5. ROUGE Score Comparison

Table 4. ROUGE scores across various feature extraction Models

Model	ROUGE-1		ROUGE-2		ROUGE-L	
	Beam	**Greedy**	**Beam**	**Greedy**	**Beam**	**Greedy**
ResNet	0.446	0.429	0.198	0.180	0.422	0.407
Inception	0.437	0.420	0.192	0.176	0.411	0.395
Xception	0.440	0.421	0.195	0.177	0.411	0.396
ViT	0.453	0.447	0.210	0.198	0.424	0.420

The ROUGE scores, which show the overlap of longer sequences between the reference and anticipated captions, are shown in Table 4. Under beam search decoding, the ViT-based model produced the highest scores for ROUGE-1, ROUGE-2, and ROUGE-L, with respective scores of 0.453, 0.210, and 0.424.

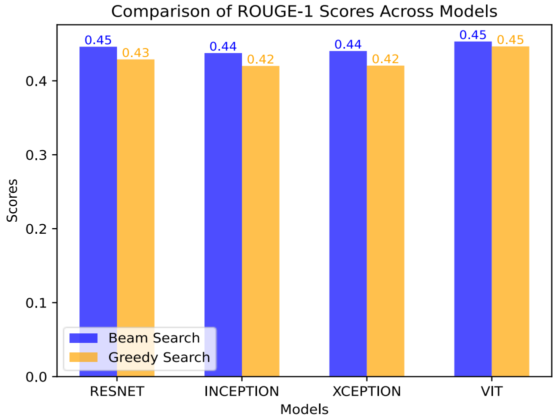

Figure 9. ROUGE-1 Scores Comparison Across Models

In figure 9 it present the ROUGE-1 scores , as it measures the overlap of individual words between generated captions and reference captions, the ViT–LSTM–LSA model records the highest score with beam search at 0.453, followed by Xception beam at 0.440, ResNet beam at 0.446, and Inception beam at 0.437. With greedy decoding, ViT achieves 0.447, staying ahead of ResNet (0.429), Xception (0.421), and Inception (0.420). The gap between beam and greedy for ViT is 0.006, smaller than that of CNN-based models, which range from 0.016 to 0.018. These results show that the ViT-based model not only leads in single-word recall but also maintains stable performance across decoding strategies, producing captions that consistently capture more of the important words from the reference descriptions.

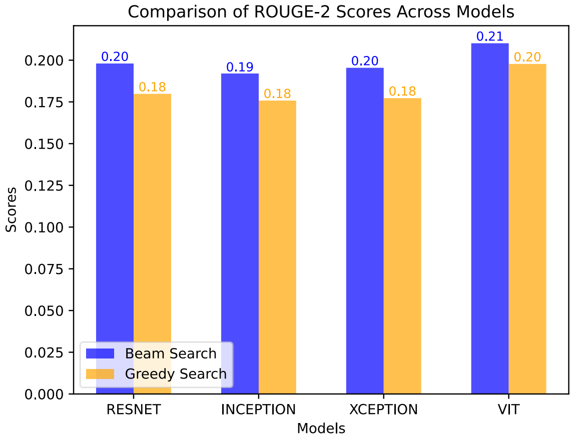

Figure 10. ROUGE-2 Scores Comparison Across Models

In figure 10 it presents ROUGE-2 scores, which measures the overlap of two-word sequences between generated and reference captions, the ViT–LSTM–LSA model achieves the highest score with beam search at 0.210, ahead of Xception beam at 0.195, ResNet beam at 0.198, and Inception beam at 0.192. In greedy decoding, ViT records 0.198, maintaining its lead over Xception (0.177), ResNet (0.180), and Inception (0.176). The beam–greedy gap for ViT is 0.012, smaller than the differences observed in the CNN-based models, which range from 0.018 to 0.020. These results highlight the model’s strong ability to preserve accurate short phrase structures and contextual meaning, while remaining consistent across both decoding strategies.

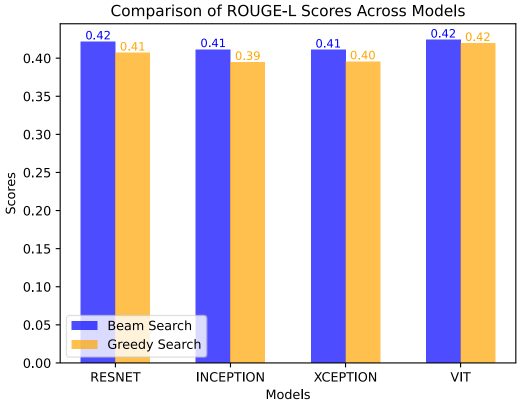

Figure 11. ROUGE-L Scores Comparison Across Models

In figure 11 ROUGE-L scores are plotted, it evaluates the longest common subsequence (LCS) between generated and reference captions and reflects overall sentence-level structure alignment, the ViT–LSTM–LSA model achieves the highest score with beam search at 0.424, followed by ResNet beam at 0.422, Xception beam at 0.411, and Inception beam at 0.411. Under greedy decoding, ViT scores 0.420, remaining ahead of ResNet (0.407), Xception (0.396), and Inception (0.395). The difference between beam and greedy for ViT is 0.004, notably smaller than the 0.014–0.016 range observed in the CNN-based models. These results indicate that the ViT-based approach is particularly effective at generating captions with sentence structures closely matching the references, while delivering stable performance across decoding methods.

The ROUGE metric results clearly show that the ViT–LSTM–LSA model consistently outperforms all CNN-based baselines in capturing both word-level and phrase-level overlaps, as well as maintaining sentence structure. It achieves the highest ROUGE-1 score of 0.453, the top ROUGE-2 score of 0.210, and the leading ROUGE-L score of 0.424 using beam search, while also retaining strong performance with greedy decoding. The small score differences between decoding methods for this model, compared to the larger gaps in CNN-based approaches, indicate its stability and reliability in generating contextually aligned captions. Overall, these results demonstrate that the ViT-based approach is more effective in preserving important content, producing fluent sentence structures, and delivering captions that closely match the meaning and flow of the reference descriptions.

5. CONCLUSION AND FUTURE SCOPE

5.1. Conclusion

In this paper, a powerful image feature extraction model is presented, which is called as Vision Transformer (ViT), which can generate captions to images by a sequence provided by the Long Short-Term Memory (LSTM), and hyperparameter optimization Lighting Search Algorithm (LSA). The proposed system utilizes the time-sequence modelling ability of LSTM and spatial context ability of ViT, and LSA enhances the model performance by adaptively varying the learning rate, dropout, and hidden units.

The proposed hybrid ViT-LSTM architecture integrates the ViT and LSTM which not only outperforms the baseline models that follow the ResNet, Inception, and Xception but also yields better results in terms of evaluation measures such as BLEU, METEOR, and ROUGE in the experimental working on flickr30k. The performance gains are attributed to the effective hyperparameter exploration strategy adopted by ViT through self-attention-based encoding as well as by LSA.

Also, the model under consideration shows the tendency to increase its effectiveness using beam search and greedy decoding strategies and have good generalization qualities. These results ratify the capacity of the proposed system in generating descriptive, fluent, and semantically true subtitles of various images.

5.2. Future Scope

Even though the proposed image captioning system has demonstrated the impressive performance on a number of benchmark datasets, there is a range of areas that can be enhanced. One positive direction is to increase the model ability to generate multiple language-specific titles to one image. This would allow the system to remember multiple contextual interpretations or semantic perspectives, something that is particularly handy in field such as accessibility, education and storytelling.

Another possible development area is the combination of structured visual information, such as scene graphs or other sources of external knowledge. The system can end up with more rational, comprehensive and human-like descriptions on combining relational and contextual knowledge of the objects in an image, especially in intricate scenes.

Real-time deployment is one more important element. Although the accuracy of the present model is optimized, it would still require further work in model compactness, quantization, or application of lightweight transformer architecture to transfer it to the edge devices or low-resource areas. This would make the system more applicable in the embedded vision applications, mobile, and surveillance.

Also, the model could be more applicable to practical applications and flexible when its implementation would be extended to multilingual or domain datasets, such as remote sensing or medical imaging. Finally, the availability of interactive learning, such as using reinforcement learning or by human feedback can allow the model to dynamically update its outputs in such a way so that the quality of the captions graduates to support the specific requirements of individual users or tasks.

6. REFERENCES

[1] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, “Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks,” *IEEE Access*, vol. 10, pp. 33679–33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[2] K. Nguyen, D. C. Bui, T. Trinh, and N. D. Vo, “EAES: Effective Augmented Embedding Spaces for Text-Based Image Captioning,” *IEEE Access*, vol. 10, pp. 32443–32452, 2022, doi: 10.1109/ACCESS.2022.3158763.

[3] H. Zhang, C. Ma, Z. Jiang, and J. Lian, “Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s,” *IEEE Access*, vol. 11, pp. 134–143, 2023, doi: 10.1109/ACCESS.2022.3232508.

[4] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with Adaptive Attention for Visual Captioning,” *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 5, pp. 1112–1131, May 2020, doi: 10.1109/TPAMI.2019.2894139.

[5] Y. Wang, N. Xu, A. A. Liu, W. Li, and Y. Zhang, "High-Order Interaction Learning for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4417–4430, Jul. 2022, doi: 10.1109/TCSVT.2021.3121062.

[6] J. Zhang, Z. Fang, H. Sun, and Z. Wang, "Adaptive Semantic-Enhanced Transformer for Image Captioning," *IEEE Trans Neural Netw Learn Syst*, vol. 35, no. 2, pp. 1785–1796, Feb. 2024, doi: 10.1109/TNNLS.2022.3185320.

[7] W. Jiang, W. Zhou, and H. Hu, "Double-Stream Position Learning Transformer Network for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7706–7718, Nov. 2022, doi: 10.1109/TCSVT.2022.3181490.

[8] L. Meng, J. Wang, R. Meng, Y. Yang, and L. Xiao, "A Multiscale Grouping Transformer with CLIP Latents for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024, doi: 10.1109/TGRS.2024.3385500.

[9] R. O. Alnashwan, S. A. Chelloug, N. S. Almalki, I. Issaoui, A. Motwakel, and A. Sayed, "Lighting Search Algorithm With Convolutional Neural Network-Based Image Captioning System for Natural Language Processing," *IEEE Access*, vol. 11, pp. 142643–142651, 2023, doi: 10.1109/ACCESS.2023.3342703.

[10] M. A. Arasi, H. M. Alshahrani, N. Alruwais, A. Motwakel, N. A. Ahmed, and A. Mohamed, "Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model," *IEEE Access*, vol. 11, pp. 104633–104642, 2023, doi: 10.1109/ACCESS.2023.3317276.

[11] D. A. Hafeth, S. Kollias, and M. Ghafoor, "Semantic Representations With Attention Networks for Boosting Image Captioning," *IEEE Access*, vol. 11, pp. 40230–40239, 2023, doi: 10.1109/ACCESS.2023.3268744.

[12] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-Quality Image Captioning with Fine-Grained and Semantic-Guided Visual Attention," *IEEE Trans Multimedia*, vol. 21, no. 7, pp. 1681–1693, Jul. 2019, doi: 10.1109/TMM.2018.2888822.

[13] I. Phueaksri, M. A. Kastner, Y. Kawanishi, T. Komamizu, and I. Ide, "An Approach to Generate a Caption for an Image Collection Using Scene Graph Generation," *IEEE Access*, vol. 11, pp. 128245–128260, 2023, doi: 10.1109/ACCESS.2023.3332098.

[14] A. Ueda, W. Yang, and K. Sugiura, "Switching Text-Based Image Encoders for Captioning Images With Text," *IEEE Access*, vol. 11, pp. 55706–55715, 2023, doi: 10.1109/ACCESS.2023.3282444.

[15] S. Chang and P. Ghamisi, "Changes to Captions: An Attentive Network for Remote Sensing Change Captioning," *IEEE Transactions on Image Processing*, vol. 32, pp. 6047–6060, 2023, doi: 10.1109/TIP.2023.3328224.

[16] Q. Wang, W. Huang, X. Zhang, and X. Li, "GLCM: Global-Local Captioning Model for Remote Sensing Image Captioning," *IEEE Trans Cybern*, vol. 53, no. 11, pp. 6910–6922, Nov. 2023, doi: 10.1109/TCYB.2022.3222606.

[17] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote Sensing Image Change Captioning With Dual-Branch Transformers: A New Method and a Large Scale Dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3218921.

[18] N. Xu *et al.*, "Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning," *IEEE Trans Multimedia*, vol. 22, no. 5, pp. 1372–1383, May 2020, doi: 10.1109/TMM.2019.2941820.

[19] J. W. Bae, S. H. Lee, W. Y. Kim, J. H. Seong, and D. H. Seo, "Image Captioning Model Using Part-of-Speech Guidance Module for Description With Diverse Vocabulary," *IEEE Access*, vol. 10, pp. 45219–45229, 2022, doi: 10.1109/ACCESS.2022.3169781.

[20] J. Zhang, K. Mei, Y. Zheng, and J. Fan, "Integrating Part of Speech Guidance for Image Captioning," *IEEE Trans Multimedia*, vol. 23, pp. 92–104, 2021, doi: 10.1109/TMM.2020.2976552.

[21] Y. Jing, X. Zhiwei, and G. Guanglai, "Context-Driven Image Caption with Global Semantic Relations of the Named Entities," *IEEE Access*, vol. 8, pp. 143584–143594, 2020, doi: 10.1109/ACCESS.2020.3013321.

[22] A. Jamil *et al.*, "Deep Learning Approaches for Image Captioning: Opportunities, Challenges and Future Potential," 2024, doi: 10.1109/ACCESS.2017.DOI.

[23] A. U. Haque, S. Ghani, and M. Saeed, "Image Captioning with Positional and Geometrical Semantics," *IEEE Access*, vol. 9, pp. 160917–160925, 2021, doi: 10.1109/ACCESS.2021.3131343.

[24] S. K. Im and K. H. Chan, "Context-Adaptive-Based Image Captioning by Bi-CARU," *IEEE Access*, vol. 11, pp. 84934–84943, 2023, doi: 10.1109/ACCESS.2023.3302512.

[25] Z. Zhou *et al.*, "An Image Captioning Model Based on Bidirectional Depth Residuals and its Application," *IEEE Access*, vol. 9, pp. 25360–25370, 2021, doi: 10.1109/ACCESS.2021.3057091.

[26] N. Ding, C. Deng, M. Tan, Q. Du, Z. Ge, and Q. Wu, "Image Captioning with Controllable and Adaptive Length Levels," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 2, pp. 764–779, Feb. 2024, doi: 10.1109/TPAMI.2023.3328298.

[27] C. Wu, S. Yuan, H. Cao, Y. Wei, and L. Wang, "Hierarchical attention-based fusion for image caption with multi-grained rewards," *IEEE Access*, vol. 8, pp. 57943–57951, 2020, doi: 10.1109/ACCESS.2020.2981513.

[28] L. Yu, J. Zhang, and Q. Wu, "Dual Attention on Pyramid Feature Maps for Image Captioning," *IEEE Trans Multimedia*, vol. 24, pp. 1775–1786, 2022, doi: 10.1109/TMM.2021.3072479.

[29] S. Chang and P. Ghamisi, "Changes to Captions: An Attentive Network for Remote Sensing Change Captioning," *IEEE Transactions on Image Processing*, vol. 32, pp. 6047–6060, 2023, doi: 10.1109/TIP.2023.3328224.

[30] J. Wang *et al.*, "Remote Sensing Image Captioning with Sequential Attention and Flexible Word Correlation,"

IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3366984.

[31] L. Zhou, Y. Zhang, Y. G. Jiang, T. Zhang, and W. Fan, "Re-Caption: Saliency-Enhanced Image Captioning through Two-Phase Learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 694–709, 2020, doi: 10.1109/TIP.2019.2928144.

[32] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global Visual Feature and Linguistic State Guided Attention for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3132095.

[33] C. Yan *et al.*, "Task-Adaptive Attention for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 43–51, Jan. 2022, doi: 10.1109/TCSVT.2021.3067449.

[34] N. Thanyawet, P. Ratsamee, Y. Uranishi, M. Kobayashi, and H. Takemura, "Identifying Disaster Regions in Images Through Attention Shifting with a Retarget Network," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3443130.

[35] T. Wei, W. Yuan, J. Luo, W. Zhang, and L. Lu, "VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning," *Journal of Systems Engineering and Electronics*, vol. 34, no. 1, pp. 9–18, Feb. 2023, doi: 10.23919/JSEE.2023.000035.