Application of Hybrid CNN-LSTM Architecture with Optuna Optimization for Weather Image Captioning

Sulaeman Salasa[1], Shintami Chusnul Hidayati[2], Muhamad Hilmil Muchtar Aditya Pradana[3]

[1,2,3]Department of Informatics, Institut Teknologi Sepuluh November, [1]Badan Meteorologi Klimatologi dan Geofisika
Email: sulaeman.salasa@bmkg.go.id

Automating the description of weather phenomena through visual imagery is a crucial step in supporting efficient meteorological monitoring systems. This study aims to compare the performance of two Deep Learning architectures, ResNet101-LSTM and VGG16-LSTM, in generating automatic image captions for various weather conditions. The research methodology involves extracting visual features using Residual Learning and VGG-Net, which are subsequently processed by Long Short-Term Memory (LSTM) units for text generation. Hyperparameter optimization was conducted using the Optuna framework to ensure both models operate at their peak configurations. The results indicate that ResNet101-LSTM provides superior linguistic accuracy, achieving a BLEU-1 score of 0.7553, a BLEU-4 score of 0.4593, and a METEOR score of 0.7264. Qualitatively, this model is capable of identifying environmental details with higher precision compared to VGG16-LSTM. However, loss curve analysis reveals that VGG16-LSTM demonstrates better convergence stability (good fit), whereas ResNet101-LSTM shows signs of slight overfitting. This study concludes that while ResNet101-LSTM is superior in accuracy according to standard NLP evaluation metrics, additional regularization techniques are required to maintain its performance stability on validation data.
Keywords: Image Captioning, Weather, ResNet101, VGG16, LSTM, Optuna.

Corresponding Author:
Sulaeman Salasa
Department of Informatics, Institut Teknologi Sepuluh November
Surabaya
sulaeman.salasa@bmkg.go.id

1. Introduction

In the meteorological ecosystem, the validity of weather data acquisition particularly in synoptic code reporting at Meteorological Stations serves as a fundamental foundation for producing accurate information. Performance evaluation studies of weather stations, both manual and automated, indicate that data validation is a critical step in minimizing observational errors [1], [2]. Weather conditions not only influence daily activities but also have crucial impacts on various sectors, ranging from the functionality of outdoor visual systems [3], structural health monitoring of buildings [4], agricultural management, to renewable energy forecasting such as solar power generation [6]. To date, operational standards for weather observation remain predominantly dominated by conventional methods that rely on manual human observation. However, Xiao et al. [26] highlight that reliance on human visual observation is not only time-consuming but also error-prone, particularly due to fatigue, physiological limitations, and observer subjectivity.

The fundamental weaknesses of conventional methods lie in their high level of subjectivity and the complexity involved in assessing atmospheric conditions. Variability in perception, disparities in expertise levels, and differences in individual experience often result in inconsistencies in recorded data, as confirmed by various instrument performance verification studies [1], [2]. In addition to accuracy issues, manual methods are also constrained by temporal limitations that create data gaps, thereby reducing information representation during extreme weather anomalies [3], [26]. These challenges are further exacerbated by

the unique characteristics of weather imagery; Elhoseiny et al. [3] explain that weather classification presents a high degree of difficulty due to complex lighting and reflection factors. Furthermore, there exists a computational challenge known as the "semantic gap" between raw visual content and human language interpretation. Various recent literature surveys [10], [11], [27] emphasize that bridging this gap requires deep image understanding to translate visual features into accurate textual descriptions.

To address these challenges, this study proposes a technical approach through the development of an Image Captioning model that integrates Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architectures. This approach adopts the Encoder–Decoder framework or the "Show and Tell" paradigm [23], which can also be implemented using memristor array designs for efficiency [5]. This hybrid architecture has proven effective in handling vision-to-language tasks across various domains, ranging from medical imaging for diagnosis [12], underwater imagery [13], and remote sensing applications [16], [25], to complex language applications such as Arabic [15] and Hindi [21]. Within this framework, the CNN functions as an encoder to extract spatial visual features from weather images in order to identify complex cloud patterns and sky conditions. The effectiveness of feature extraction can be further enhanced using spatially aware architectures [14].

The reliability of CNNs in extracting detailed visual features has been demonstrated across numerous domains, ranging from general object classification using Very Deep Convolutional Networks (VGGNet) [33] or Deep Residual Learning (ResNet) [28], to specific applications such as medical diagnosis of diabetic retinopathy [31] and plant disease identification [30]. Recent studies by Kaur and Kaur [20] as well as Kumar et al. (as cited in [20]) further indicate that efficient CNN–LSTM–based frameworks are a key factor in improving image captioning performance.

The extracted visual features are subsequently processed by the LSTM, which acts as a decoder. The selection of LSTM is based on its advantages over standard Recurrent Neural Networks (RNNs) in addressing the vanishing gradient problem and learning long-term dependencies [32]. Various architectural developments, such as the use of Stack Parallel LSTM [17], graph-based captioning [7], and long-range sequence modeling approaches such as Mamba [8], demonstrate the evolution of sequential data processing. Mujawar and Iyer [9] explain that deep learning models with coordinated relationships enable more precise sentence construction. In addition, visual attention mechanisms [22], including dual attention [18] and attention-augmented residual approaches [19], are often incorporated to focus the model on salient objects.

In this study, the CNN–LSTM architecture is employed to generate standardized weather descriptions. The evaluation of the generated descriptions is not solely based on precision but also on semantic correlation with human judgment, measured using automatic metrics such as METEOR [24]. Finally, to ensure optimal model convergence, advanced hyperparameter optimization frameworks such as Optuna [29] can be applied during the training process. Through this implementation, the study aims to automate weather classification in order to produce reliable, consistent, and human subjectivity–free observational data.

2. Method

Research Design

This research employs a mixed-methods approach, integrating quantitative and qualitative analysis. The quantitative approach is used to numerically measure model performance using standard evaluation metrics, while the qualitative approach is applied to analyze the semantic coherence of the generated weather descriptions and visualize word distribution. The research workflow is systematically designed,

starting from data acquisition and re-captioning, model development, parameter optimization, and performance evaluation.

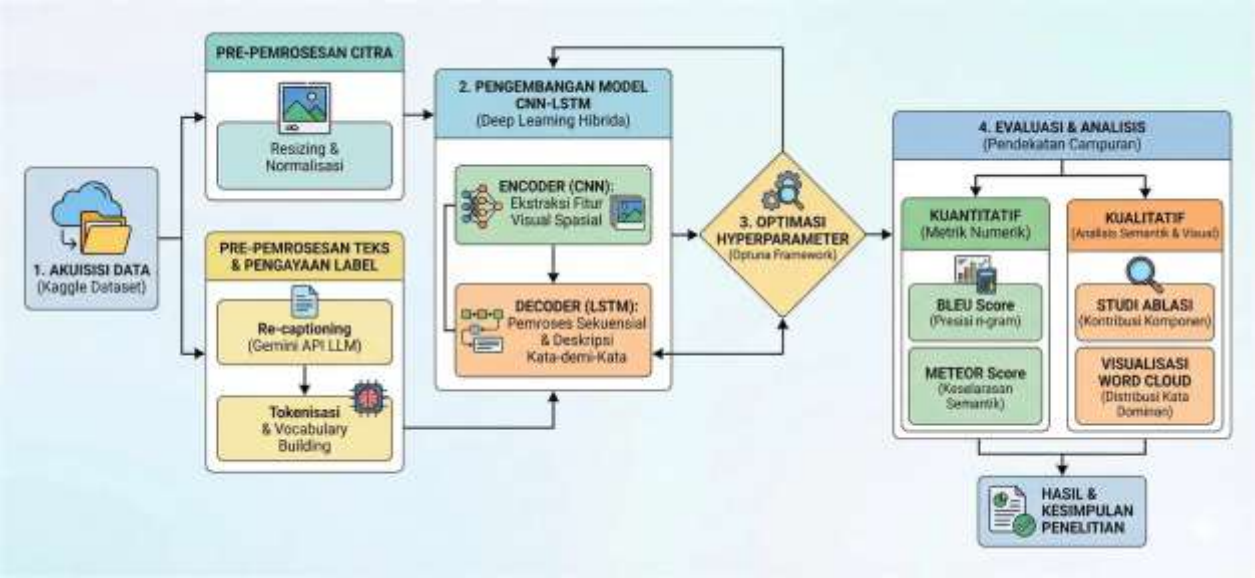

Figure 1. Research Flowchart

The research process begins with Data Acquisition, where datasets are collected from Kaggle sources. This raw data is then processed through two separate pathways. The first pathway is Image Preprocessing, which includes Resizing and Normalization techniques to standardize image formats. The second pathway is Text Preprocessing & Label Enrichment, which involves Re-captioning using the Gemini API LLM to improve descriptions, and Tokenization and Vocabulary Building to convert text into numeric data. The results of these two preprocessing steps serve as input for the model development phase.

The next phase is 2. CNN-LSTM Model Development, a Hybrid Deep Learning architecture. This model consists of two main components: an Encoder (CNN) that extracts spatial visual features from images, and a Decoder (LSTM) as a sequential processor that generates verbatim descriptions. To achieve optimal performance, 3. Hyperparameter Optimization is performed using the Optuna Framework, which iteratively adjusts the CNN-LSTM model parameters.

After the model was developed and optimized, a mixed-methods approach was used. Quantitative evaluation used numerical metrics such as the BLEU Score to measure n-gram precision and the METEOR Score to assess semantic alignment. Meanwhile, qualitative analysis included an Ablation Study to understand the contribution of model components and a Word Cloud Visualization to observe the distribution of dominant words. All findings from this evaluation stage were then summarized into the Research Results & Conclusions.

Model Architecture: CNN-LSTM

This system architecture is built on the foundation of a synergistic hybrid model between CNN and LSTM. The first component, the Encoder, uses a CNN to extract visual features. The CNN processes the input weather image and produces feature maps rich in spatial information. These feature maps represent the essential visual characteristics of the image, including elements such as cloud formation and texture, sky color gradation, and captured lighting conditions. The CNN's ability to capture these complex visual patterns is crucial as it provides the foundation for subsequent processing stages.

Once the visual features are extracted, the second component, the Decoder, takes over the processing role. This decoder uses an LSTM architecture specifically designed to handle sequential data. The LSTM receives

feature vectors generated by a CNN and processes them to understand the context and relationships between these visual elements. The LSTM's primary advantage lies in its ability to learn long-term or temporal dependencies between words. In this process, the LSTM constructs the weather description incrementally, predicting word-by-word based on the received visual features and previously generated words, until a complete, coherent, and syntactically accurate weather description is formed. This synergy between the CNN's spatial feature extraction capabilities and the LSTM's sequential processing enables the system to generate relevant textual descriptions from weather images.

Hyperparameter Optimization with Optuna

To achieve optimal model performance, this study implemented an automatic hyperparameter search mechanism using the Optuna framework. Optuna was used to efficiently explore a wide parameter space, including settings for the learning rate, number of hidden units in the LSTM, batch size, and dropout rate. This process aims to balance the model's ability to learn patterns (fitting) while preventing overfitting, so that the model can adapt well to the validation data.

Table 1. Hypertuning Parameter Table

Parameter	Baseline Model (VGG-16 + LSTM)	Proposed Model (CNN-LSTM)
Architecture	VGG-16 (CNN) + LSTM	ResNet-101 (CNN) + LSTM
Training Strategy	Full Fine-Tuning (Pre-trained)	Full Fine-Tuning (Pre-trained)
Optimizer	Adam	Adam
Learning Rate	3×10^{-4}	5×10^{-4}
Batch Size	16	32
Epochs	30	30
Regularization	Dropout (0.5)	Dropout (0.5)
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy

Data Evaluation and Analysis

The testing and analysis phase is conducted through several mechanisms to ensure the validity of the results. The first mechanism is a quantitative evaluation using industry-standard metrics, namely BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering). The BLEU metric is used to measure n-gram precision, which reflects how accurately the model reproduces phrases contained in the reference sentence. The BLEU score is calculated using the following equation:

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

In this formulation, the variable N represents the total number of n-grams used in the evaluation. Each n-gram level contributes to the model, governed by a weight wn, which is generally uniformly equal to $1/N$. The Pn component refers to the precision of the n-gram, modified to avoid overfitting. Furthermore, this calculation takes into account sentence length through the Brevity Penalty (BP), which depends on c (the length of the predicted/candidate sentence) and r (the length of the effective reference sentence).

$$F_{mean} = \frac{10.P.R}{R + 9.P}$$

$$pen = \gamma.\left(\frac{ch}{m}\right)$$

$$M = (1 - Pen).F_{mean}$$

Meanwhile, the METEOR metric is used to measure the semantic alignment between model predictions and the reference, taking into account precision and recall. The METEOR score is calculated using the equation The variable Fmean is the harmonic mean score combining the precision P and recall R values, with greater

emphasis on recall. The variable P is Unigram Precision, which is the ratio of matched unigrams to the total unigrams in the candidate. The variable R is Unigram Recall, which is the ratio of matched unigrams to the total unigrams in the reference. The Penalty component is calculated based on the fragmentation of word matches, where ch (the number of chunks) is the sequentially matched chunks of a word sequence, and m is the total number of matched unigrams. The parameters γ and θ (which influence the Penalty) are determined empirically.

In addition to the metric evaluation, an ablation study was also conducted to analyze the specific contribution of each model component. This analysis was performed by modifying or removing certain components, such as variations in CNN layers or LSTM configurations, to understand their impact on overall performance. Finally, a qualitative analysis was conducted using word cloud visualization to map the frequency distribution of the most frequently occurring words in the model predictions. This visualization provides insight into the model's descriptive focus on dominant weather elements and validates the relevance of the learned vocabulary to the meteorological domain.

3. Result and Discussion

Quantitative Analysis

A quantitative evaluation was conducted to measure the extent to which the CNN-LSTM model was able to learn visual patterns from the weather dataset and translate them into accurate text descriptions. The quantitative analysis method used was the loss curve analysis of the CNN-LSTM model training. The following method was used to compare the values of the Bleu-1, Bleu-4, and METEOR metrics to measure the resulting word order and the meaningfulness of the CNN-LSTM model.

1. ResNet101-LSTM Convergence Performance
   The training process was conducted for 100 epochs for each model. The main evaluation parameter used was the loss function to monitor how well the model predicted word order in weather descriptions. This approach aligns with the Show and Tell methodology developed by Vinyals et al. [23], where log-likelihood optimization of the target description is performed to generate accurate sentences.
   Based on Figure 3, a visualization of the Training Loss versus Validation Loss graph, the ResNet101-LSTM model shows a very sharp decrease in training loss, reaching 0.23. This is possible because the Residual Learning architecture [28] in ResNet101 is able to overcome the problem of vanishing gradients in very deep networks, allowing the model to extract more complex visual features than conventional CNN architectures.
   However, the validation loss graph confirms the presence of overfitting symptoms. As noted in the literature on deep learning for weather [26], overly complex models risk "memorizing" the noise in the training dataset if not accompanied by appropriate regularization techniques. This phenomenon is clearly visible from the widening gap between the training and validation curves after the 20th epoch.
2. VGG16-LSTM Convergence Performance
   In contrast to ResNet, the VGG16-LSTM model exhibits a healthier convergence pattern (good fit). Although the final training loss value is higher (0.38), the agreement between the training and validation curves indicates superior generalization ability.
   The use of VGG16 as a feature extractor remains relevant due to its uniform structure using 3 * 3 filters [33], which has been shown to be effective in recognizing fundamental visual features. This combination with Long Short-Term Memory (LSTM) allows the model to remember word sequence dependencies in a stable manner over the long term, in accordance with the basic principles introduced by Hochreiter & Schmidhuber [32].

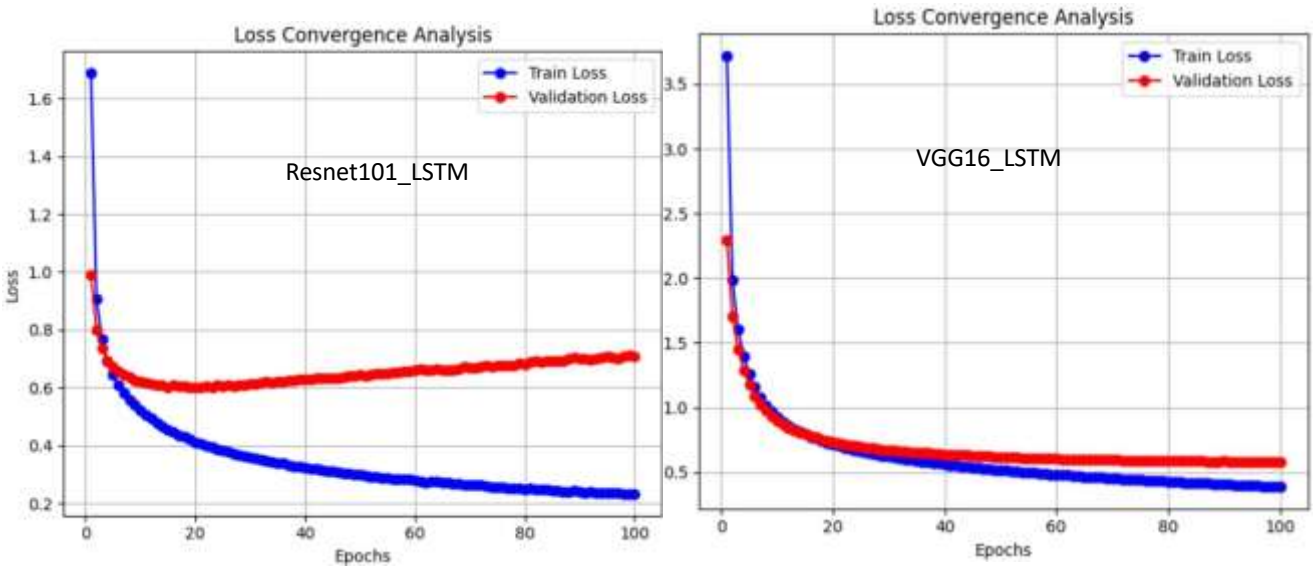

Figure 2. Loss curves for CNN-LSTM (ResNet101 vs. VGG16)

3. Comparison of loss metrics

These results indicate that in the context of weather image description, a deeper architecture (ResNet101) does not always guarantee better results on validation data compared to a simpler but more stable architecture (VGG16). This finding supports the review by Zhang et al. [25], which states that model selection for remote sensing and meteorology should consider a balance between network depth and dataset capacity to avoid performance degradation on unfamiliar data.

Table 2. presents a summary of the final performance of both models:

Model Architecture	Final Training Loss	Final Validation Loss	Convergence Status	Architecture Reference
ResNet101–LSTM	0.23	0.70	Overfitting	He et al. [28]
VGG16–LSTM	0.38	0.58	Good Fit	Simonyan & Zisserman [33]

4. Hyperparameter Optimization Analysis with Optuna

To achieve optimal model performance, this study implemented an automatic hyperparameter search mechanism using the Optuna framework. This process is crucial for balancing the model's ability to learn patterns (fitting) while preventing overfitting. Based on the experimental results summarized in Table 1.1, the proposed ResNet101-LSTM model was configured with a learning rate of 5 * 10-4 and a batch size of 32.

The slightly higher learning rate setting for ResNet101 compared to the baseline VGG16 model (3 * 10-4) indicates that the Residual Learning architecture is capable of handling more aggressive gradient updates without losing stability early in the training phase. Although Optuna optimized the regularization parameter with a dropout rate of 0.5 for both models, ResNet101-LSTM still exhibited signs of overfitting in the convergence graph. This proves that in very deep architectures, hyperparameter optimization alone is not enough to reduce model complexity for the weather dataset used, so a balance between network depth and dataset capacity is still needed to avoid performance degradation on unknown data.

Table 3. Hyperparameter Optimization Analysis with Optuna

Parameter	Baseline Model (VGG-16 + LSTM)	Proposed Model (ResNet-101 + LSTM)	Impact Analysis
Learning Rate	3×10^{-4}	5×10^{-4}	ResNet uses a slightly higher learning rate, indicating that the Residual Learning architecture can handle more aggressive gradient updates without losing stability in the early training stage[4].
Batch Size	16	32	The use of a larger batch size in ResNet-101 helps stabilize gradient estimation, which is crucial for very deep networks[5].
Regularization (Dropout)	0.5	0.5	Both models apply a relatively high dropout rate (0.5) to mitigate the risk of overfitting[6].

5. Model Performance Analysis Based on NLP Metrics

After training and validation, the performance of both models was evaluated using standard Natural Language Processing (NLP) metrics, namely BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering). These metrics are crucial for measuring the similarity between machine-generated weather descriptions and human references, with the BLEU score used to measure the precision of the generated text's n-grams. Based on the test results, the ResNet101-LSTM model performed significantly better on the BLEU-1 (Unigram) metric, with a score of 0.7553 compared to the VGG16-LSTM model, which scored 0.6788. This difference of 0.0765 indicates that the ResNet101's Residual Learning architecture is more accurate in extracting visual features to select individual vocabulary words, such as nouns or weather attributes. This superiority trend continued in the BLEU-4 (4-gram) score, where ResNet101-LSTM led with a score of 0.4593, a 0.0421-point advantage over VGG16-LSTM. This indicates that the sentence or phrase structure generated by the ResNet-based model is more coherent and closer to the patterns of human reference language, in line with the advantages of the CNN-LSTM hybrid in handling temporal data sequences.

Unlike BLEU, which focuses on exact word matches, the METEOR metric places weight on stemming and synonyms, resulting in a higher correlation with human judgment. The data shows that the highest improvement occurred in this metric, with a difference of +0.0946, with ResNet101-LSTM achieving a score of 0.7264 and VGG16-LSTM 0.6318. This superiority reinforces the indication that the visual features extracted by ResNet101 provide richer semantic information for the LSTM decoder to generate more natural and varied sentences compared to the VGG16 architecture. Overall, although optimization using the Optuna framework was applied to both models to find the most efficient hyperparameters, ResNet101-LSTM consistently demonstrated better linguistic performance in visually describing weather phenomena. This finding also confirms that using more complex models like ResNet can provide better spatial representation for automated meteorological applications than traditional sequential models.

Table 4. BLEU and Meteor Scores

Evaluation Metrics	Model CNN-LSTM (Resnet 101)	*Model CNN-LSTM (VGG-16)*	Difference (Improvement)
BLEU-1	0.7553	*0.6788*	+0,0765
BLEU-4	0.4593	0.4172	+0.0421 (Resnet 101 Unggul)

Evaluation Metrics	Model CNN-LSTM (Resnet 101)	*Model CNN-LSTM (VGG-16)*	Difference (Improvement)
METEOR	0.7264	0.6318	+0.0946 (Resnet 101 Unggul)

Qualitative Analysis

Qualitative analysis was conducted to determine the model's ability to translate visual features captured by the Vision Transformer into meteorologically plausible descriptive sentences. This testing was conducted by sampling data from a test set that the model had not previously encountered during the training process.

1. Comparison of Caption Generation Results

Table 5. Comparison of VGG16 vs. Resnet 101

Weather Image	Ground Truth Caption	Predicted Caption
	CNN(VGG-16)-LSTM	
	The sky is covered with dust with limited visibility, showing a dust storm phenomenon that significantly reduces visibility. This condition affects outdoor activities and transportation.	The sky is covered with dust with very dry air conditions, showing a dust storm phenomenon that significantly reduces visibility. This condition has the potential to disrupt daily activities and respiratory health.
	Clear sky with trees covered in frost, showing a frost phenomenon that creates a beautiful winter landscape. This condition can affect road conditions and visibility.	Clear sky with frost covering the trees, showing a significant frost phenomenon. This condition causes reduced visibility and an extremely cold environment.
	The sky is covered with thick clouds with heavy snowfall, showing snow that reduces visibility and covers road surfaces. This condition may cause disruptions to transportation and outdoor activities.	The sky is covered with thick clouds with heavy snowfall, showing snow that reduces visibility and covers road surfaces. This condition may cause disruptions to transportation and outdoor activities.
	CNN(RESNET-101)-LSTM	
	Clear sky without clouds, showing frost on trees indicating extremely cold temperatures. This condition can affect road conditions and the surrounding environment.	Clear sky without clouds, showing frost covering the trees. This condition can affect visibility and cause slippery roads.
	Clear sky without clouds, showing snow covering the trees. Accumulated snow can affect environmental and vegetation conditions.	Clear sky without clouds, showing snow covering trees and the ground. This condition causes reduced visibility and a cold environment.

Weather Image	Ground Truth Caption	Predicted Caption
	CNN(VGG-16)-LSTM	
	The sky is covered with thick clouds with snowfall. Snow covers the ground surface and trees, significantly reducing visibility.	The sky is covered with thick clouds with heavy snowfall, causing limited visibility and an environment covered with a layer of snow. Heavy snowfall affects outdoor activities.

2. Visual Phenomenon Identification Accuracy
   The ResNet101-LSTM model demonstrated deeper feature extraction capabilities in recognizing specific elements in the image. For example, in the image of snow-covered trees against a blue sky, the model was able to accurately identify the condition "clear sky without clouds." Conversely, in the dust storm sample, the VGG16-LSTM model tended to provide a more general description of "very dry air conditions" while still successfully capturing the essence of the main phenomenon, such as "dust storm."
3. Sentence Structure and Contextual Detail
   Both models were able to produce coherent sentences that adhered to the subject-verb-object structure well, reflecting the effectiveness of using LSTM in modeling language sequences. However, there were differences in the details of weather impacts:
   a. ResNet101-LSTM: Tends to provide technical details regarding environmental impacts, such as mentioning "the roads became slippery" or "the environment became cold" due to snow accumulation.
   b. VGG16-LSTM: Focuses on the broad impacts of activities, such as mentioning potential disruptions to "respiratory health" in the context of dust storms or "daily activities."
4. Consistency with Ground Truth (Original Caption)
   Visually, the descriptions generated by ResNet101-LSTM have very high semantic similarity to the original captions, especially in the object identification sections such as "heavy snowfall" and "limited visibility." This is supported by the high METEOR score (0.7264), indicating that the model's synonym selection and word stemming closely approximate how humans describe weather.
5. Error and Hallucination Analysis
   While overall performance was good, slight differences were found in the interpretation of object textures. In the frost image, the VGG16-LSTM model described "frost cover on trees" very closely to the original caption, while ResNet101-LSTM provided additional descriptions of "soil" that may not be explicitly visible in the image's primary focus. This relates to the findings in the previous loss graph, where the complexity of ResNet sometimes causes the model to be too specific in describing details (a symptom of mild overfitting at the visual level).

4. Conclusion

Based on the research results and discussion, it can be concluded that the implementation of the ResNet101-LSTM and VGG16-LSTM architectures for weather image captioning systems exhibits significantly different characteristics. Quantitatively, the ResNet101-LSTM model provides superior linguistic accuracy performance compared to VGG16-LSTM. This is evidenced by the BLEU-1 score of 0.7553, BLEU-4 of 0.4593, and METEOR of 0.7264. The superiority of the METEOR score by a difference of +0.0946 confirms that the use of Residual Learning is capable of extracting richer visual features, resulting in more natural sentences with strong semantic closeness to human references. Qualitatively, visual analysis shows that ResNet101-LSTM is more detailed in identifying environmental elements, such

as snow texture and cloudless skies, while VGG16-LSTM tends to produce more functional and general descriptions. However, in terms of training stability, the VGG16-

LSTM model exhibited a healthier convergence profile (good fit) with a stable validation loss of 0.58. In contrast, ResNet101-LSTM indicated mild overfitting despite being optimized using the Optuna framework. This suggests a trade-off where the deeper architecture (ResNet101) excels at capturing visual details for high accuracy, while the simpler architecture (VGG16) offers better generalization stability. Overall, ResNet101-LSTM is recommended for applications that prioritize description precision, while VGG16-LSTM is more reliable for systems that require consistency across varying data.

5. References

[1] H. Subyantara Wicaksana et al., "Evaluasi Kinerja Automatic Weather Station Berdasarkan Pengamatan Paralel di Stasiun Meteorologi Kemayoran," 2021.

[2] D. R. Wibawanty, W. Wandayantolis, and I. Ishak, "Verifikasi Kinerja Alat Automatic Weather System (AWS) dan Termometer Digital terhadap Observasi Manual di Stasiun Klimatologi Palembang," *JRST (Jurnal Riset Sains dan Teknologi)*, vol. 6, no. 2, p. 151, Nov. 2022, doi: 10.30595/jrst.v6i2.13541.

[3] M. Elhoseiny, S. Huang, and A. Elgamma, "Weather classification with deep convolutional neural networks," in *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, Canada, 2015. doi: 10.1109/ICIP.2015.7351424.

[4] N. N. H. Dinh, H. Shin, Y. Ahn, B. L. Oo, and B. T. H. Lim, "Attention-based image captioning for structural health assessment of apartment buildings," *Autom. Constr.*, vol. 167, Nov. 2024, doi: 10.1016/j.autcon.2024.105677.

[5] Y. Yu et al., "Neural image caption generator based on crossbar array design of memristor module," *Neurocompu- ting*, vol. 560, Dec. 2023, doi: 10.1016/j.neucom.2023.126766.

[6] M. Abumohsen, A. Y. Owda, M. Owda, and A. Abumihsan, "Hybrid machine learning model combining of CNN- LSTM-RF for time series forecasting of Solar Power Generation," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 9, Sep. 2024, doi: 10.1016/j.prime.2024.100636.

[7] M. J. Parseh and S. Ghadiri, "Graph-based image captioning with semantic and spatial features," *Signal Process. Image Commun.*, vol. 133, Apr. 2025, doi: 10.1016/j.image.2025.117273.

[8] T. Shahzad, M. Aoun, T. Mazhar, M. U. Tariq, K. Ouahada, and H. Hamam, "Mamba-caption: Long-range se- quence modelling for efficient and accurate image captioning," *Array*, vol. 28, Dec. 2025, doi: 10.1016/j.ar- ray.2025.100538.

[9] S. R. Mujawar and S. Iyer, "Deep learning model with co-ordinated relationship for image captioning enabled via attentional language encoder-decoder," *Signal Process. Image Commun.*, vol. 142, p. 117466, Mar. 2026, doi: 10.1016/j.image.2025.117466.

[10] A. Sharma, H. Singh, and M. Pant, "Pixels to prose: A comprehensive survey of image captioning techniques with deep learning and generative artificial intelligence," *Neurocomputing*, Feb. 28, 2026, doi: 10.1016/j.neu- com.2025.132385.

[11] H. D. Abdulgalil and O. A. Basir, "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models," *Natural Language Processing Journal*, Sep. 01, 2025, doi: 10.1016/j.nlp.2025.100159.

[12] F. Firdaus et al., "A medical image captioning system for TeleOTIVA: Supporting SDGs-oriented cervical pre- cancer screening in Indonesia," *Inform. Med. Unlocked*, vol. 60, p. 101719, Jan. 2026, doi: 10.1016/j.imu.2025.101719.

[13] L. Li, H. Li, and P. Ren, "Underwater image captioning via attention mechanism based fusion of visual

and textual information," *Information Fusion*, vol. 123, Nov. 2025, doi: 10.1016/j.inffus.2025.103269.
[14] V. P. Saxena et al., "Enhancing Image Understanding in Automatic Captioning using Spatially-Aware Transformer Architectures," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 2081–2090. doi: 10.1016/j.procs.2025.04.458.
[15] N. Aljojo, H. Ardah, A. Tashkandi, and S. Habibullah, "Predicting abnormality-guided multimodal linguistic se- mantics Arabic image captioning," *Machine Learning with Applications*, vol. 21, p. 100706, Sep. 2025, doi: 10.1016/j.mlwa.2025.100706.
[16] M. R. Sree, M. Siddhartha, P. V. Vardhan Reddy, B. Kruthika, and R. P. Singh, "A Residual Network and Bi-direc- tional LSTM based Hybrid Approach to Remote Sensing Image Captioning," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 88–97. doi: 10.1016/j.procs.2025.04.198.
[17] X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, and X. Niu, "Image captioning with triple-attention and stack parallel LSTM," *Neurocomputing*, vol. 319, pp. 55–65, Nov. 2018, doi: 10.1016/j.neucom.2018.08.069.
[18] R. Padate, A. Jain, M. Kalla, and A. Sharma, "Image caption generation using a dual attention mechanism," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106112.
[19] Mrs. A. P and Dr. P. D, "Image Captioning System for Natural Language Processing using Optimized Attention- Augmented Residual Convolutional Neural Network," *Knowl. Based. Syst.*, p. 115272, Jan. 2026, doi: 10.1016/j.knosys.2026.115272.
[20] M. Kaur and H. Kaur, "An Efficient CNN-LSTM Based Framework for Improved Image Captioning," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 3601–3607. doi: 10.1016/j.procs.2025.04.615.
[21] A. K. Poddar and R. Rani, "Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 686–696. doi: 10.1016/j.procs.2023.01.049.
[22] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.
[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[24] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop*, 2005.
[25] K. Zhang, P. Li, and J. Wang, "A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Mod- els, Comparisons and Future Directions," *Remote Sensing*, vol. 16, no. 21, 4113, Nov. 2024, doi: 10.3390/rs16214113.
[26] H. Xiao, F. Zhang, Z. Shen, K. Wu, and J. Zhang, "Classification of Weather Phenomenon From Images by Using Deep Convolutional Neural Network," *Earth and Space Science*, vol. 8, no. 5, May 2021, doi: 10.1029/2020EA001604.
[27] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, "A thorough review of models, evaluation metrics, and datasets on image captioning," *IET Image Process.*, vol. 16, pp. 311–332, Feb. 2022, doi: 10.1049/ipr2.12367.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[29] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.
[30] S. Mulyana et al., "Identifikasi Penyakit Tanaman Berdasarkan Citra Daun Berbasis Web dengan Pendekatan Algoritma Convolutional Neural Network," *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 8, no. 2, pp. 305–317, 2025.
[31] E. Hari Rachmawanto and M. Muslih, "Convolutional Neural Network (CNN) untuk Klasifikasi Citra

Penyakit Diabetes Retinopathy," *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 5, no. 2, pp. 167–176, 2022.

[32] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Interna- tional Conference on Learning Representations*, 2015.