12th INTERNATIONAL SCIENTIFIC CONFERENCE ON DEFENSIVE TECHNOLOGIES
OTEH 2026
Serbia, Tara, 21 - 24 October 2026

AUTOMATED IMAGE CAPTIONING WITH A RESNET-LSTM ENCODER-DECODER FOR DISTRIBUTED SURVEILLANCE SCENES

ANĐA DENIĆ
Faculty of Electronic Engineering, University of Niš, Niš, andja.denic@elfak.ni.ac.rs

SNEŽANA ZUROVAC
Military Technical Institute, Belgrade, zurovac@medianis.net

IVANA PETKOVIĆ
Military Technical Institute, Belgrade, ivana.petkovic.vti@gmail.com

Abstract: *Automatic image captioning is a fundamental artificial intelligence task at the intersection of computer vision and natural language processing. This paper presents a generative encoder-decoder model for automatic caption generation from images of military exercises. The model employs an ImageNet-pre-trained 152-layer Residual Network (ResNet-152) encoder to extract visual representations and a long short-term memory (LSTM) decoder to generate corresponding textual descriptions. The original model was trained on the Microsoft Common Objects in Context (MS COCO) dataset and subsequently fine-tuned on a domain-specific collection of military-exercise images. Five pseudo-reference captions per image were generated automatically using Florence-2 and used to supervise domain adaptation.*

Experimental evaluation was performed on a held-out military-domain test set containing 25 images. The fine-tuned model achieved Bilingual Evaluation Understudy scores of 70.87, 49.67, 33.61, and 21.94 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively, together with a Recall-Oriented Understudy for Gisting Evaluation based on the longest common subsequence (ROUGE-L) score of 45.52. On the same military test set, the original MS COCO checkpoint achieved 52.00, 33.10, 23.00, 15.80, and 37.05, respectively. These results indicate that domain-specific fine-tuning improves agreement with the available military-domain reference captions.

However, because the test set is small and its reference captions were generated automatically rather than independently written by domain experts, the reported results should be interpreted as preliminary evidence rather than a definitive estimate of performance on the broader military-image domain. The proposed architecture may provide a basis for future investigation of onboard semantic reporting and bandwidth-efficient communication on unmanned platforms, while its practical contribution can be further experimentally validated through the evaluation of execution speed, energy consumption, and communication savings on a real unmanned aerial vehicle.

Keywords: *image captioning, LSTM, ResNet, encoder-decoder, domain-specific dataset, military-exercise*

1. INTRODUCTION

Automatic image captioning addresses one of the central challenges in machine learning: converting visual perception into structured natural language. Unlike image classification or object detection, where the output is usually a fixed label or a set of localized categories, caption generation requires a model to identify objects, infer their attributes and relations, and express the resulting scene understanding as a grammatically coherent sentence. This makes image captioning a representative vision-language problem at the intersection of computer vision and natural language processing.

The practical relevance of automatic image captioning is particularly strong in scenarios where large volumes of visual data must be interpreted and communicated efficiently. In surveillance, reconnaissance, digital archiving, assistive technologies, and autonomous robotic systems, raw image and video streams can contain substantially more data than can be transmitted, stored, or inspected in real time. A compact textual description may provide a semantic summary of a scene and support filtering, indexing, prioritization, and operator decision-making. For distributed systems, including unmanned platforms, onboard caption generation could potentially reduce the amount of visual data that must be transmitted to a central station. Nevertheless, the present study evaluates only captioning accuracy on an offline image dataset. It does not experimentally evaluate onboard inference, real-time processing, energy consumption, communication latency, or operation on an actual unmanned platform.

In this paper, we study automatic caption generation for images of military exercises. This domain introduces a meaningful distribution shift with respect to general-purpose image-captioning datasets because the scenes may contain specialized equipment, formations, activities, uniforms, vehicles, and operational contexts that are underrepresented in broad collections of everyday

Corresponding author: A. Denić;
E-mail: andja.denic@elfak.ni.ac.rs
DOI: 10.5937/OTEH26066D

imagery. The original model was trained on the Microsoft Common Objects in Context (MS COCO) dataset, which provides useful general visual and linguistic regularities but does not adequately represent all concepts found in military-training environments. Domain adaptation is therefore investigated as a means of improving the relevance of generated descriptions for the target imagery.

The proposed approach follows an encoder-decoder paradigm. A 152-layer Residual Network (ResNet-152) acts as the visual encoder and transforms an input image into a fixed-dimensional representation. A long short-term memory (LSTM) network then generates the caption sequentially, one word at a time. The model is first trained on MS COCO and subsequently fine-tuned on automatically annotated military-exercise images. The principal contribution is an end-to-end domain-adaptation pipeline that connects target-domain data collection, automatic annotation, deterministic image preprocessing, image-level dataset partitioning, parameter-efficient fine-tuning, and comparative evaluation of the original and adapted checkpoints.

This architecture offers a favorable balance between modeling capacity and deployment practicality. While large vision-language foundation models have recently achieved impressive performance, compact encoder-decoder systems remain attractive for edge scenarios where computational resources, latency, power consumption, and communication bandwidth are constrained. In such settings, a ResNet-LSTM model can provide interpretable and efficient caption generation without requiring continuous transmission of high-resolution visual streams.

The main contribution of this work is an end-to-end captioning pipeline adapted to military-exercise imagery. The paper demonstrates how a general image-captioning model can be transferred from MS COCO to a specialized operational domain, evaluates the impact of fine-tuning using standard captioning metrics, and discusses the suitability of the resulting system for distributed surveillance and drone-based deployment. Experimental results indicate that domain-specific fine-tuning improves the quality and relevance of generated captions compared with the base model trained only on generic data.

2. THEORETICAL BACKGROUND AND RELATED WORK

The present study extends the encoder-decoder framework examined in the preceding work. A central instance of this framework is the neural image-caption generator introduced by Vinyals et al. [1], in which a convolutional neural network (CNN) encodes an image and a recurrent neural network (RNN) estimates the probability of a caption conditioned on the resulting visual representation.

Let I denote an input image and Y its caption of length T. The model factorizes the conditional probability of the complete caption autoregressively as

$$P_\theta(Y \mid I) = \prod_{t=1}^{T} P_\theta(y_t \mid I, y_1, y_2, \ldots, y_{t-1}) \qquad (1)$$

where θ contains the trainable parameters of the visual encoder, word-embedding layer, recurrent decoder, and output classifier. Learning is performed by maximizing the conditional likelihood of the reference captions, or equivalently by minimizing token-level cross-entropy. This probabilistic formulation provides a direct connection between visual representation learning and sequence modeling and is the theoretical basis of both the original model and its domain-specific adaptation considered here.

The visual encoder used in this work is based on ResNet-152. Residual networks introduce identity shortcuts that allow a stack of nonlinear layers to learn a residual mapping rather than an unconstrained direct mapping, thereby facilitating the optimization of very deep convolutional architectures [2]. In the adopted implementation, the classification layer of a ResNet-152 model pre-trained on ImageNet is removed. Global image features produced by the convolutional backbone are transformed by an additional linear layer and batch-normalization layer into a fixed-dimensional context vector. Transfer learning is particularly important in the present setting because a comparatively small collection of military-exercise images is insufficient for learning a deep visual representation from random initialization.

The decoder is a long short-term memory network. LSTM units were designed to model long-range dependencies through a recurrent cell state regulated by input, forget, and output gates [3]. For image captioning, the encoded image is supplied to the recurrent model before the textual sequence, after which the decoder predicts one word at a time. The probability distribution over the vocabulary at time step t depends on the current LSTM hidden state, which summarizes the image representation and the previously observed or generated words. During training, reference tokens are provided as preceding inputs, whereas inference is autoregressive and uses the model's own previous predictions.

MS COCO dataset provides a broad source domain for learning general visual and linguistic regularities because it contains complex everyday scenes paired with multiple human-written captions [4]. Nevertheless, its content distribution differs from that of military-training imagery. Uniforms, formations, specialized vehicles, aircraft, vessels, weapons, field exercises, and operational activities may occur rarely or under different contextual priors in a general-purpose dataset. As a result, a model trained only on MS COCO can produce grammatically plausible captions while omitting domain-relevant details or selecting vocabulary that is semantically inappropriate for the target environment. The problem addressed in this paper is therefore one of domain adaptation: knowledge acquired from a large generic dataset is retained, while the trainable components are updated using target-domain image-caption pairs.

The broader development of vision-language modeling has introduced attention mechanisms, Transformer architectures, and large prompt-conditioned foundation models. Visual attention allows a decoder to select spatially relevant image regions while generating

individual words [5], whereas Transformer-based models replace recurrence with attention-based sequence processing [6]. Florence-2 represents a more recent unified approach: it treats captioning, detection, grounding, and related tasks as prompt-conditioned sequence generation within a single vision-language model [7]. In the present pipeline, Florence-2 is not the final deployed captioning model and its parameters are not fine-tuned. Instead, the pre-trained Florence-2 model is used as an automatic annotation system that generates pseudo-reference captions for the military-exercise images. These annotations subsequently supervise the fine-tuning of the more compact ResNet-LSTM model.

Domain-specific captioning has also been investigated for remote-sensing imagery, where the distribution of viewpoints, spatial structures, and semantic categories differs substantially from ordinary ground-level photography [8]. This line of work illustrates the general principle relevant to defence-oriented data: a captioning architecture may be transferable, but its vocabulary and conditional output distribution must be adapted to the target domain. For surveillance and drone-based platforms, compact encoder-decoder systems remain attractive because they can convert high-bandwidth visual observations into short semantic reports locally. Such use does not eliminate the need for human verification, particularly in safety-critical or operational contexts, but it can support filtering, indexing, prioritization, and bandwidth-efficient communication.

3. METHODOLOGY AND DATA

3.1. Dataset

The target-domain dataset was constructed from publicly available photographs of military exercises published on the official website of the Ministry of Defence of the Republic of Serbia. The images depict military personnel, formations, training procedures, helicopters, boats, ground vehicles, equipment, and outdoor exercise environments that are not adequately represented in general-purpose captioning datasets. The complete collection contains 167 distinct images. Of these, 142 images are used for training and validation, while 25 images form a held-out military-domain test set.

Because manually producing several independent descriptions for every image is time-consuming, the images were automatically annotated using the open-source Florence-2 model, specifically microsoft/Florence-2-base-ft. Annotation was performed using a graphics processing unit (GPU), and five pseudo-reference captions were generated for each image. These captions provide scalable supervision for fine-tuning the smaller ResNet-LSTM model, but they should not be treated as equivalent to independently written expert annotations.

The quality of the automatically generated captions directly affects both model training and evaluation. Florence-2 may omit relevant objects or actions, use overly general terminology, repeat similar sentence structures, or occasionally introduce an object that is not present in the image. Such errors act as label noise during fine-tuning and can encourage the adapted model to reproduce the same omissions, biases, or hallucinations. Moreover, because the evaluation references were generated by the same automatic annotation procedure, BLEU and ROUGE-L primarily measure agreement with Florence-2 wording and may not fully reflect factual correctness or agreement with a military-domain expert. Five captions produced by one model are also likely to be linguistically correlated and therefore do not provide the same diversity as five descriptions written independently by human annotators.

The training-validation split is performed at the image level rather than at the image-caption-pair level. Unique image identifiers are sorted, shuffled using a fixed pseudorandom seed, and partitioned using a validation ratio of 0.15. This produces 121 training images with 605 captions and 21 validation images with 105 captions. Because all captions belonging to one image remain in the same subset, direct image leakage between training and validation data is prevented. The independent test collection contains 25 images and 125 automatically generated reference captions and is used to compare the original and fine-tuned checkpoints.

The limited size of the target dataset is an important limitation of the study. In particular, results calculated on only 25 test images can be sensitive to the presence or absence of a small number of scenes, objects, and environmental conditions. Consequently, the reported metric improvements may not generalize to the full diversity of military imagery, including different exercises, equipment categories, seasons, weather conditions, illumination levels, viewpoints, sensors, and geographic environments. The current evaluation should therefore be regarded as an initial controlled comparison on the available test collection. A larger independently collected test set, preferably with captions reviewed or written by domain experts, is required for a more reliable estimate of generalization.

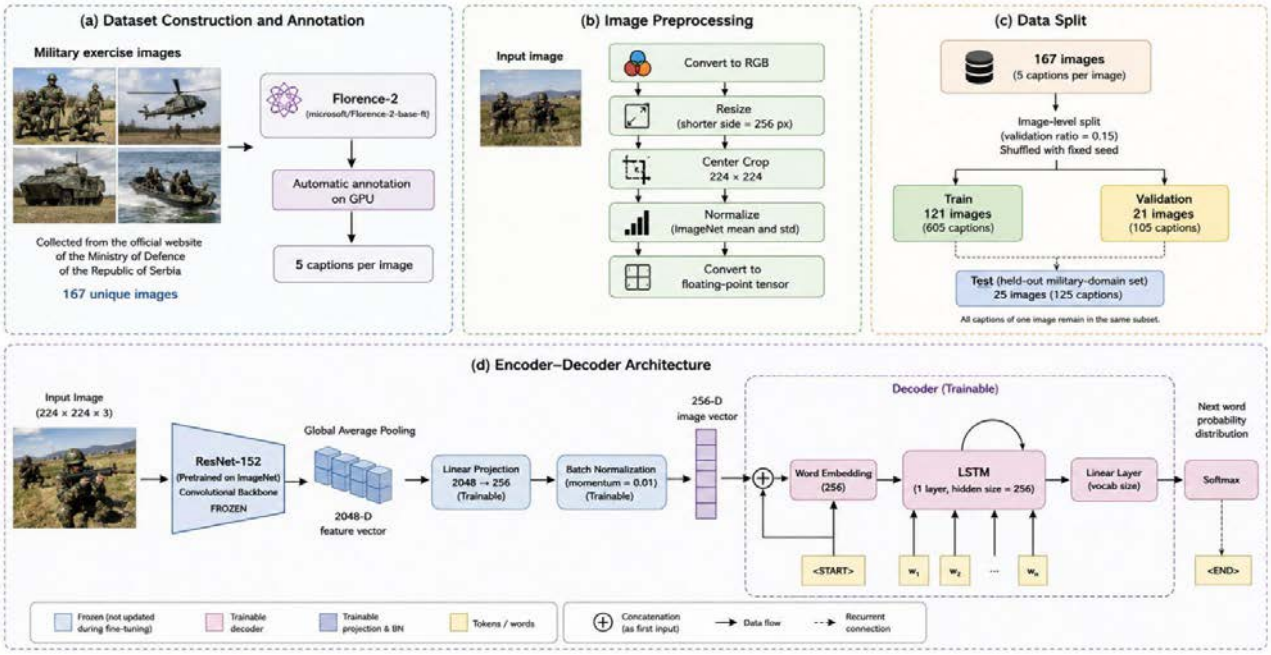


Figure 1. Overview of the methodology: (a) dataset construction and automatic annotation with Florence-2; (b) image preprocessing pipeline; (c) train/validation/test split performed at the image level; (d) encoder-decoder architecture with a frozen ResNet-152 encoder and trainable LSTM decoder

Figure 1 presents an overview of the complete methodology used in this study, from dataset construction and preprocessing to model training and inference. Before being passed to the ResNet encoder, each image is opened with the Python Imaging Library, converted to RGB, resized so that the shorter side equals 256 pixels, center-cropped to 224×224 pixels, and converted to a floating-point tensor. Channel-wise normalization uses the ImageNet statistics μ=(0.485, 0.456, 0.406) and σ=(0.229, 0.224, 0.225), which is consistent with the pre-training of the ResNet backbone. The same deterministic transformation is used during fine-tuning and evaluation.

Because manually producing several independent captions for every image is costly, the images were automatically annotated using the open-source Florence-2 model (microsoft/Florence-2-base-ft). Annotation is performed on the GPU. Five captions have been collected for an image.

Fine-tuning performed the training-validation split at the image level rather than at the image-caption-pair level. Unique image identifiers are sorted, shuffled using a fixed pseudorandom seed, and partitioned using a validation ratio of 0.15. This yields 121 training images and 21 validation images. Because all captions associated with one image remain in the same subset, the split prevents direct image leakage between training and validation data. The held-out military-domain test collection is reserved for the comparison between the original and fine-tuned models.

3.2. Encoder-Decoder Architecture

The target model retains the architecture developed in the preceding study [1]. Its encoder is an ImageNet-pre-trained ResNet-152 with the original classification layer removed. After global average pooling, the convolutional backbone produces a 2,048-dimensional vector. A trainable linear projection maps this vector to a 256-dimensional joint representation, followed by batch normalization with momentum 0.01. All parameters in the ResNet convolutional backbone are frozen, whereas the projection and batch-normalization parameters remain trainable.

The decoder contains a word-embedding layer of dimension 256, a single-layer LSTM with a hidden-state dimension of 256, and a linear output layer whose dimension equals the vocabulary size. The image vector and word embeddings consequently occupy the same dimensional space. Following the Show and Tell formulation [1], the encoded image is inserted as the first input to the LSTM, before the <START> token and caption words. The final linear layer transforms each LSTM hidden state into unnormalized vocabulary scores, and a softmax interpretation yields the conditional probability of the next token.

Captions are converted to lowercase and punctuation is removed. The sequence is tokenized by whitespace, mapped to indices from the vocabulary inherited from the MS COCO model, and surrounded by <START> and <END> markers. Variable-length sequences are sorted within each mini-batch, padded, and packed for efficient recurrent processing. Tokens absent from the inherited vocabulary are excluded by the current preprocessing implementation rather than being used as explicit <UNK> targets. This choice preserves compatibility with the pre-trained decoder checkpoint but may remove specialized domain terminology; extending the vocabulary is therefore an important direction for subsequent work.

3.3. Training on MS COCO

The base model parameters are those obtained through training on MS COCO 2017. The source-domain training set contains more than 118,000 images, and the validation set contains 5,000 images, with approximately five human-written captions per image. The model uses a 256-dimensional image and word representation, a 256-dimensional LSTM state, and one recurrent layer. Training is conducted for 10 epochs with mini-batches of 64 examples and the Adam optimizer with a learning rate of 10^{-3}.

For a mini-batch, the encoder first produces image feature vectors, and the decoder generates vocabulary scores under teacher forcing.

The trainable parameters, denoted as θ, are the solutions to an optimization problem that maximizes the conditional likelihood of the reference captions given the input images.

The optimizer updates the complete decoder together with the encoder's linear projection and batch-normalization layer; the ResNet-152 convolutional backbone remains fixed. At the end of each epoch, the average validation loss is computed without gradient tracking. The encoder and decoder checkpoints associated with the lowest validation loss are retained for inference and evaluation. At inference time, the model uses greedy decoding: the most probable word is selected at each step until the <END> token is produced or a maximum caption length of 20 tokens is reached.

3.4. Domain-Specific Fine-Tuning

Fine-tuning starts from the saved encoder and decoder checkpoints of the MS COCO model; no component is reinitialized. The same Encoder CNN and Decoder RNN classes, vocabulary, image transformation, loss definition, and optimizer family are used to ensure architectural compatibility with the source model. The target-domain training and validation loaders use a batch size of 64.

The domain-adaptation stage is run for five epochs with Adam and a learning rate of 10^{-3}. Consequently, the frozen ResNet backbone preserves general ImageNet visual features, while the projection and language decoder adapt the joint image-text representation and conditional word distribution to the military-exercise domain. This parameter-efficient strategy reduces the computational cost and the risk of overfitting relative to updating all ResNet-152 layers on the limited target dataset.

Each epoch consists of a training phase and a validation phase. Whenever the validation loss improves, the fine-tuned state dictionaries overwrite the previous encoder and decoder checkpoints.

4. EXPERIMENTAL RESULTS

The experiments were conducted using an NVIDIA Quadro RTX 4000 graphics processing unit (GPU) with 8 GB of video memory, an eight-core central processing unit (CPU) supporting 16 processing threads, and 64 GB of random-access memory (RAM). The implementation was developed in Python using PyTorch. Torchvision provides the pre-trained ResNet-152 architecture and image transformations, while Florence-2 annotations are generated through the Hugging Face Transformers interface.

Caption quality is evaluated using the Bilingual Evaluation Understudy (BLEU) metric from BLEU-1 through BLEU-4 and the Recall-Oriented Understudy for Gisting Evaluation based on the longest common subsequence (ROUGE-L). BLEU measures modified n-gram precision with a brevity penalty [9], with larger n placing progressively greater emphasis on multiword correspondence. ROUGE-L measures sequence-level overlap using the longest common subsequence and therefore considers both shared content and word order [10]. All values in Table 1 are reported as percentages.

Table 1. Captioning performance on the base and military-domain evaluation sets (%)

Model / evaluation domain	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Base model / MS COCO	68.36	50.35	35.70	25.23	49.51
Original model / military test set	52.00	33.10	23.00	15.80	37.05
Fine-tuned model / military test set	**70.87**	**49.67**	**33.61**	**21.94**	**45.52**

When the original MS COCO checkpoint is applied directly to the military test set, all metrics decrease relative to its source-domain performance, confirming the presence of a distribution shift. Fine-tuning improves BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-L on the same military test set by 18.87, 16.57, 10.61, 6.14, and 8.47 percentage points, respectively. Because both checkpoints are evaluated using the same images and reference captions, this comparison provides evidence that adaptation improves agreement with the available target-domain annotations.

Nevertheless, the military test set contains only 25 images. Metric values calculated from such a small collection may have high sampling variability and can be disproportionately influenced by a few easy or difficult examples. No confidence intervals or repeated evaluations on independently sampled military test sets are available. The observed improvements are therefore valid for the evaluated collection but should not be interpreted as a precise estimate of performance across the broader military-image domain.

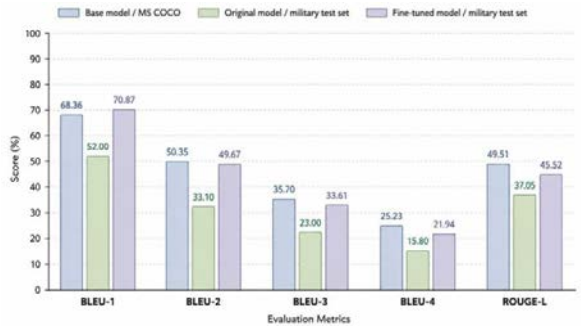


Figure 2. Captioning performance on the base (MS COCO) and military-domain evaluation sets. The fine-tuned model outperforms the original model on all metrics of the military test set

Figure 2 provides a visual comparison of the captioning performance achieved by the base and fine-tuned models on the MS COCO and military-domain evaluation datasets using the BLEU and ROUGE-L metrics. The scores obtained by the base model on MS COCO are included to document its source-domain performance, but they should not be interpreted as a controlled comparison with the military-domain scores because the reference captions and image distributions differ. The relevant evidence for the effectiveness of fine-tuning is the comparison of the two checkpoints on the same military test set.



Figure 3. Comparison of captions generated by the original model and the fine-tuned model for the first representative image from the military test set

Original model: “a man is sitting on a chair with a laptop”
Fine-tuned model: “two soldiers are sitting inside of a military vehicle”



Figure 4. Comparison of captions generated by the original model and the fine-tuned model for the second representative image from the military test set

Original model: “a group of people standing around a field of grass”

Fine-tuned model: “a group of soldiers in military uniforms are standing in a line”



Figure 5. Comparison of captions generated by the original model and the fine-tuned model for the second representative image from the military test set

Original model: “a group of people standing near a fire”

Fine-tuned model: “a group of soldiers walking through a forest”



Figure 6. Comparison of captions generated by the original model and the fine-tuned model for the second representative image from the military test set

Original model: “people standing beside a yellow truck”

Fine-tuned model: “a group of soldiers standing next to a tank”

Figures 3–6 illustrate both the benefits and the remaining limitations of domain-specific fine-tuning. In Figure 3, the original model focuses on familiar indoor objects and produces the caption “a man is sitting on a chair with a laptop,” whereas the fine-tuned model identifies the military context and describes two soldiers inside a military vehicle. Similarly, in Figure 4, the original model generates the overly general description “a group of people standing around a field of grass,” while the fine-tuned model more accurately identifies soldiers in military uniforms standing in a line.

The examples also reveal several characteristic errors. For the low-light firefighting scene in Figure 5, the caption “a group of people standing near a fire” is broadly related to the image but incomplete: it does not identify the people as military firefighters, does not describe their protective

equipment, and omits the active response to a forest fire.

Figure 6 demonstrates that domain adaptation can improve one part of a caption while introducing a different error. The original model describes the scene as “people standing beside a yellow truck,” which recognizes the presence of people and a large yellow vehicle but does not identify either the military personnel or the tracked construction machine. The fine-tuned model generates “a group of soldiers standing next to a tank.” This caption correctly introduces the military personnel but incorrectly classifies the tracked bulldozer as a tank. The error suggests that the adapted model may over-apply military-domain priors when a visually similar but non-combat vehicle appears in a military setting.

The observed qualitative errors can therefore be grouped into three categories: overly general captions that omit domain-relevant information, incomplete captions that identify the main scene but omit important objects or actions, and incorrect object classifications caused by visual similarity or target-domain bias. These examples also show why improvements in BLEU and ROUGE-L should be complemented by human evaluation of factual correctness. Since the qualitative analysis contains only a small number of selected images, it illustrates possible failure modes but does not establish their frequency in the broader target domain.

5. CONCLUSION

This paper extends a ResNet-152-LSTM image-captioning model trained on MS COCO to a specialized collection of military-exercise imagery. The principal contribution is an end-to-end domain-adaptation pipeline that combines automatic target-domain annotation with parameter-efficient fine-tuning. A pre-trained Florence-2 model generates five pseudo-reference captions per image, after which the original encoder-decoder checkpoint is adapted by updating the complete LSTM decoder and the encoder's projection and batch-normalization layers while keeping the convolutional ResNet backbone fixed.

Evaluation on military-domain images demonstrates a clear advantage over direct application of the source model. Fine-tuning improves BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-L by 18.87, 16.57, 10.61, 6.14, and 8.47 percentage points, respectively.

Beyond improving the quality of the generated captions, the results of this study also indicate the significant potential of such models for applications in modern defence technologies. During reconnaissance missions using unmanned platforms, video streams are typically transmitted continuously from the onboard camera to the ground control station. For example, transmitting a Full HD video stream (1920 × 1080 pixels, 30 fps) compressed using the H.264 codec at a bitrate of 5 Mb/s requires the transmission of approximately 625 kB of data per second. In contrast, a textual description of a single image generated by the proposed model occupies approximately 150 bytes. When processing 30 images per second, this corresponds to a data rate of only about 4.5 kB/s. These results open up the possibility of significantly reducing the amount of data transmitted over the communication link during the communication between the unmanned aerial vehicle and the ground control station, which may be particularly important under conditions of limited bandwidth.

Additional reductions in the required communication bandwidth can be achieved by applying an event-driven information transmission strategy. In this case, the system would continuously analyse images directly on the unmanned platform, while textual descriptions would be transmitted only if the generated caption contains predefined objects or activities of interest, such as military personnel, combat vehicles, helicopters, vessels, or other assets. This approach avoids transmitting a large number of irrelevant scenes, further reducing the communication load and improving the overall efficiency of the system.

Although this study focuses on automatic image captioning, the proposed approach also opens the possibility for the development of semantic communication systems, in which only information with operational significance is transmitted instead of the complete video stream. Such a concept could represent an important direction for the development of future intelligent reconnaissance systems, particularly in scenarios with limited communication bandwidth and the need for timely decision-making.

Several extensions should be considered in future work. First, the dataset should be expanded across a broader range of exercises, seasons, weather conditions, viewpoints, platforms, and equipment categories, with strict source-level partitioning to reduce correlations between training and test scenes. Second, a subset of automatically generated captions should be independently reviewed or rewritten by domain experts, and inter-annotator agreement should be reported. Finally, because defence-oriented captioning is safety-relevant, future work should quantify uncertainty, identify hallucinated objects or activities, and define a human-in-the-loop protocol under which generated descriptions support rather than replace operator judgment.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

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

[4] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Computer Vision - ECCV 2014, pp. 740-755, 2014.

[5] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proc. 32nd Int. Conf. Machine Learning (ICML), 2015, pp. 2048-2057.

[6] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] B. Xiao et al., "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 4818-4829.

[8] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2183-2195, 2018.