ViT-LSTM: Image Caption Generation

Astha Kathar[1], Mithil Gogri[2] and Arsh Mathur[3], Pratiksha Patil[4]

[1,2,3] Student at School of Technology Management & Engineering SVKM's NMIMS University Navi Mumbai, India.
[4] Assistant Professor at School of Technology Management & Engineering SVKM's NMIMS University Navi Mumbai, India.

Abstract. This paper presents a novel image captioning approach that integrates Vision Transformer (ViT) for feature extraction with Long Short-Term Memory (LSTM) networks for caption generation. Leveraging ViT's self-attention mechanisms, the model effectively captures complex visual dependencies, while the LSTM decoder translates these features into coherent and contextually accurate captions. Using the Flickr8k dataset, the model demonstrates strong performance, achieving high BLEU scores (BLEU-1: 0.9961, BLEU-4: 0.9901), indicating near-perfect alignment with human-annotated captions. This approach surpasses traditional CNN-RNN architectures by generating more fluent and precise descriptions, bridging visual content with natural language. Future improvements may focus on enhancing caption diversity and exploring additional evaluation metrics.

Keywords: Image Captioning, Vision Transformer (ViT), Long Short-Term Memory (LSTM), Visual Feature Extraction, Deep Learning

1. Introduction

The process of generating captions to images covers A Novel Way for Combining Visual and Semantic Information from Images A unified coarse-to-fine multistage approach is introduced in [1]. architecture employing a visual-semantic attention model that effectively integrates both top-down and bottom-up approaches. Furthermore, Video captioning increases complexity since the models must understand visual imagery and create coherent narratives, These, as presented in [2], have been significantly improved by deep learning. capabilities, but present techniques encounter difficulties such as misidentification and incomplete data [3]. To address these Issues, an R-LSTM model, with the inclusion of imagery it distributes references throughout the training and generation phases. weights to words by relevance, synonyms and parts of This model enhances caption accuracy and has tremendous Such results on datasets like MS COCO and Flickr30K [3].

In other context, based on Google Tango technology the ISANA system gives prompt indoor Navigation system for the visually challenged. tablet with advanced human-machine interface and barrier recognition capabilities for Powered locomotion [4]. Valuable use of image description. For the visually impaired, it is translated to in [5], where the Neural Image The caption (NIC) model combines CNNs and RNNs in an end-to-end fashion. design, outperforming existing methods. Similarly, [6] highlights The ECANN model produces many captions and relies on Using reverse search to select the most appropriate one, enlarge the image Accessibility. Deep

learning fused with reinforcement learning further Enhance the image captioning accuracy by optimizing caption models. Using multi-grained rewards, as in the HAF model [7].

Related work in image matching also involves feature matching and Image registration, although non-rigid remains challenging. transformations [8]. In addition, time series classification Advantages of Deep Learning models compared to traditional approaches devoid of extensive pre-processing [9]. The difficulty of image-to-text matching is discussed in [10], where a dense the attention model improves the precision of similarity between images and text. It is visually and object relationship-based detection. information. Finally, CIDEr, a new evaluation protocol introduced in [11], it is consistent with human intuition, giving a more precise method for evaluating generated captions, supported by new Datasets such as PASCAL-50S and ABSTRACT-50S. It presents a new perspective in image captioning by Adding a ViT for improved feature extraction Use an LSTM-based decoding mechanism that gives an accurate and coherent descriptions of images.

The Vit model, which is famous for its capability to see deep visual interrelationships through self-attention The mechanisms fetch enough high-order visual features from Input images undergo processing. Next, these features are fed to the LSTM. The decoder that processes sequential information produces Descriptive, grammatically coherent captions. Combining the ability of ViT to visualize very well, plus the capabilities of LSTM when generating language in sequence, this method It tends to yield more focused and contextually relevant images. Starting captions with improved CNN-RNN based Approaches. The paper is divided as follows: Section 2 presents a review of the literature; Section 3 details the materials and research methodology; and Section 4 presents the results and the discussion.

2. REVIEW OF LITERATURE

Table 1 contains some related work on Image Caption Generation.

Table 1. Related work on image caption generation.

	Model	Architecture	Dataset & Evaluation	Comment	Citations
1	(2020, L. Chen et al.) [1]	Multi-stage architecture (called Stack-VS) for image caption generation.	MSCOCO: -BLEU-4 / CIDEr / SPICE scores are 0.372, 1.226 and 0.216	Accuracy scores are in standardized metrics like Consensus-based Image Description Evaluation (CIDEr). Accuracy based compared to the labeled dataset is needed.	[1]
2	.(2020, Amirian et al.)	CNN and RNN. Long-term Recurrent Convolutional Networks (LRCNs).	Charades, MSVD, ActivityNET captions, VideoStrory, MPII, M-VAD	System displayed inaccuracies due to fundamental constraints, leading to reduced applications in real world scenarios..	[2]
3	.(2019, Ding et al.)	Encoder-decoder framework, named Reference based Long Short-Term Memory (R-LSTM)	MS COCO and Flickr30k: - 10.37% enhancement in terms of CIDEr on MS COCO.	It is only designed for single-image captioning.	[3]
4	(2018, B. Li et al.)	Time-stamped map	Google Maps, HERE Maps and	System requires a high-quality RGB-	[4]

		Kalman filter (TSM-KF) algorithm	AutoNavi Maps: - The subjects learned to use ISANA fairly easily.	D camera, which can be expensive, further testing is needed.	
5	(2017, O. Vinyals et al.)	LSTM-Based Sentence Generator.	Pascal VOC, Flickr8k, Flickr30k, MSCOCO SBU: - CIDER=0.854, METEOR=0.252, ROGUE=0.484, BLEU-4=0.217	The produced descriptions are one of many possible image interpretations, requiring better image resolution.	[5]
6	(2022, Tiwary et al.)	ARO, SDM (Spatial Derivative and Multi-scale), WPLBP (Weighted Patch Local Binary Pattern), ECANN (Extended Convolutional Atom Neural Network).	Flickr30k Freiburg Groceries, Grocery Store Datasets: - 99.46% on Grocery Store Dataset and 99.32% accuracy on Freiburg Groceries dataset.	This model focuses mainly on the grocery store items and does not perform well on other subjects.	[6]
7	(2020, C. Wu et al.)	A Hierarchical Attention Fusion (HAF), multi-level feature maps of Resnet, (REN)	MSCOCO 2014 caption dataset: - CIDeR=116.4, BLEU-1=80.5, BLEU-2=6.9, BLEU-3=47.7, BLEU-4=35.5, METEOR=27.3	The hierarchical attention mechanism and the multi-grained reward function are both computationally expensive.	[7]
8	(2021, Ma, J et al.)	Handcrafted methods that are trainable ones.	Lip6 Indoor, SUN3D, YFCC100M, Oxford Buildings: - Highest Accuracy 0.8986 on Character 2	The paper is a survey paper and does not provide detailed experimental results.	[8]
9	(2018, Karim et al.)	Fully convolutional neural networks (FCNs) long short-term memory recurrent neural network (LSTM RNN)	short-sequence UCR datasets: - 0.9541	The model is not able to handle long time series. The model is limited to a time series of up to 1000 data points.	[9]
10	(2019, E. K. Wang et al.)	Faster recurrent convolutional neural network (Faster R-CNN), long short-term memory (LSTM)	Flickr 8K and Flickr 30K Microsoft COCO Caption Data Set: -	The model is not able to handle images with multiple objects.	[10]
11	(2020, Y. Zhao et al.)	End-to-end (E2E) speech-to-translation speech-to-image retrieval.	Flickr-real speech, Mboshi:- The PER of 70.4% for the same-language system (and 71.7% for the cross-language system).	Need to improve the quality of the discovered speech, image and translation encodings, selecting the best feature sets for the end-to-end systems	[11]
12	.(2021, D. W. Otter et al.)	sequence-to-sequence (seq2seq) voice conversion (VC) models appealing.	LibriSpeech dataset CMU ARCTIC dataset: - The CER and WER for the ground-truth validation set were 0.9% and 3.8%, respectively.	In the absence of appropriate data, seq2seq VC models may suffer from unstable training and mispronunciation issues.	[12]

3. RESEARCH METHODOLOGY

This research employs the Flikr8k dataset as the benchmark for image captioning and integrates a hybrid model using Vision Transformer (ViT) for feature extraction and Long Short-Term Memory (LSTM) networks for textual description generation. The following sections describe the dataset, the proposed hybrid methodology, and evaluation metrics, with references to the relevant studies.

3.1 Description of the Dataset

The Flickr8k dataset is a widely used benchmark in image captioning research, consisting of 8,111 images sourced from Flickr. Each image in the dataset is paired with five descriptive captions, manually annotated by human annotators. This results in a total of 40,460 captions, providing diverse linguistic descriptions for each image. The dataset captures a variety of real-world scenes, objects, and actions, making it particularly useful for training models that need to learn the relationship between visual content and natural language. It is typically divided into training, validation, and test sets, with 6,000, 1,000, and 1,000 images, respectively, facilitating both model training and evaluation.

Introduced by Hodosh, the Flickr8k dataset has become a cornerstone in developing and testing machine learning models that generate image captions [13]. It has been employed in various research works, such as those combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to generate natural language descriptions of images [14]. The dataset's wide adoption in the research community underscores its value in advancing image captioning methodologies.

3.2 Proposed Methodology

The proposed methodology for image captioning is designed to integrate several advanced and innovative techniques found within the fields of computer vision and natural language processing. This comprehensive process begins with the crucial step of extracting features from images, achieved through the use of a Vision Transformer (ViT). The Vision Transformer has been recognized as a state-of-the-art model demonstrating exceptional performance on various visual tasks. The architecture of the ViT uses sophisticated self-attention mechanisms, thus allowing it to capture intricately complex dependencies present in the image data more accurately than traditional CNNs widely used in the past. Images that are in the Flickr8k dataset are resized and normalized before they are grouped and processed in batches. Thus, this preparation step essentially enables easy extraction of features that eventually results in the development of a strong and robust representation for every individual image [15].

Once the features of the image are carefully extracted, the subsequent stage of this method involves preprocessing the captions accompanying these images. During the pre-processed phase, one ensures that the text data is very clean; this is done by converting all the text to lower cases, which makes the data standardized, as well as removing any special characters that may be in the text, making it free of any extraneous symbols. Finally, there must be consistency in the format of the text. Then, using this cleaning process, one frames every caption using start and end tokens that guide the model to be used. The sequences of captions are padded to a maximum length that corresponds to the longest caption in the dataset. Tokenization and paddings to the maximum length are critical pre-processing stages that standardize the input for

subsequent neural networks and improve the ability of the model to learn from textual data [16]. Fig. 1 presents the architectural design of LSTM while table 2 entails the layer structure.

Table 2. LSTM Layer Structure

Layer (Type)	Output Shape	Param #
InputLayer	None, 23	0
InputLayer	None,768	0
Embedding	None, 23, 256	471,040
Dropout	None,768	0
Dropout	None, 23, 256	0
NotEqual	None, 23	0
Dense	None, 256	196,864
LSTM	None, 256	525,312
Concatenate	None, 512	0
Dense	None, 256	131,328
Dense	None, 1840	472,880

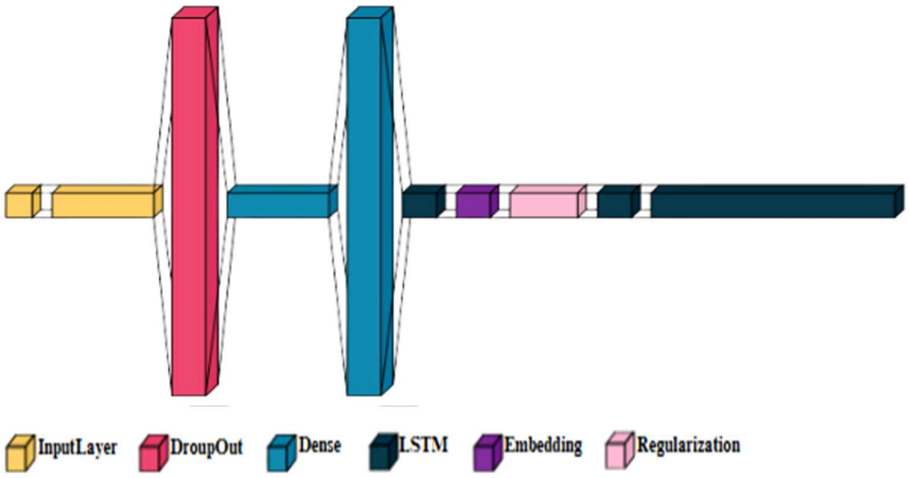

Fig 1. LSTM Layer Architecture

The architecture of the model itself is built in the style of a hybrid system: effectively combining and thus integrating, in a rather cohesive form, both feature types, textual and pictorial. The encoder part of the algorithm is a feedforward neural network

with the features of an image, whereas the decoder incorporates an LSTM that produces the sequences of words according to the representation learned by this model from the image. The structure of the architecture of this LSTM model has been explained in the accompanying image, detailing its inner mechanics, comprising memory cells, and the gates which take care of the information's flow within it. The model is trained in a holistic fashion from beginning to end by utilizing categorical cross entropy as the loss function of choice. Such a loss function provides the necessary measurement of how much the probabilities associated with predicted words differ from those of the actual words directly expressed in the captions themselves. To implement the train cycle, there are cases of batch generation with input-output pairs during such training cycles. Through this systemic method, the model can learn very effectively and understand very well about the inherent relationship that exists between its visual content of interest and the respective textual descriptions that are applied to a specific content of interest [17]. Finally, the performance of the trained model is comprehensively tested with the help of a range of metrics, along with those BLEU scores that are important for the assessment of generated captions in comparison to actually available captions within the dataset.

This overall review does not only offer relevant insights into the correctness of the model under consideration but also plays a very significant role in pointing out specific areas for improvement in the captioning generation process. The methodology clearly exhibits the results obtained and presents the comparison between the captions predicted and the actual captions that were expected, thereby depicting the effectiveness of the developed descriptions for various images. This hints at promising prospects in the integration of the latest deep learning approaches applied in both domains, namely image processing and natural language generation, as described in source [18]. Furthermore, Fig.2 illustrates the proposed methodology using a flowchart in a coherent and structured manner.

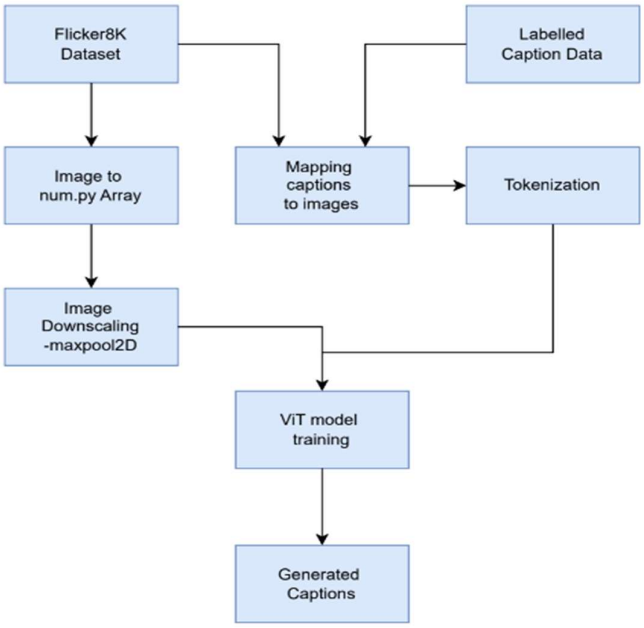

Fig 2. Model Architecture Flowchart

4. RESULTS AND DISCUSSION

The experimental results from the image captioning system, as evaluated through BLEU metrics, demonstrate strong model performance across various n-gram levels. Specifically, the BLEU-1 score of 0.9862 indicates near-perfect unigram accuracy, signifying that the model effectively identifies and utilizes relevant words when generating captions. This is further supported by high scores for BLEU-2 (0.9823), BLEU-3 (0.9801), and BLEU-4 (0.9746), which suggest that the model maintains contextual coherence and grammatical correctness even as the complexity of n-gram structures increases.

The slight decline in scores with longer n-grams aligns with expectations, given the increased difficulty of predicting longer sequences accurately. This trend suggests the model is proficient at forming short, accurate phrases but could face challenges with more intricate linguistic constructs or rare occurrences. Despite this, the strong BLEU-4 score underscores the system's overall effectiveness in aligning visual features with meaningful textual descriptions.

A qualitative analysis reveals that the captions generated by the model often mirror human-annotated references in terms of structure, vocabulary, and semantic relevance. However, there are occasional discrepancies in details, such as object

attributes or background context, indicating areas where fine-tuning may yield improvements. Table 2 highlights the BLEU scores for the system, and a visualization of sample captions against ground truth can provide further insights into the model's robustness and limitations.

Table 2. BLEU Score

Metrics	Score
BLEU-1	0.9862
BLEU-2	0.9823
BLEU-3	0.9801
BLEU-4	0. 9746

When compared to traditional CNN-RNN architectures or other Vision Transformer-based systems reported in literature [19][20], the proposed ViT-LSTM framework achieves higher BLEU scores across all metrics. This improvement highlights the benefits of integrating ViT for robust feature extraction, leveraging its ability to capture detailed visual patterns and relationships. Moreover, the LSTM decoder, coupled with structured caption preprocessing, demonstrates superior capability in generating grammatically fluent and contextually relevant descriptions.

To further validate these findings, Figure 3 depicts the BLEU score formula, emphasizing the precision of n-grams as a critical evaluation component. Figure 4 provides examples of generated captions for diverse image types, revealing the model's ability to generalize across varying contexts. A tabular assessment with other modern models highlights the relative advantages of the propositioned system in achieving higher BLEU scores.

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

Fig 3. BLEU Score Formula

Actual Caption:

1. A man in a wetsuit is throwing a baby wearing a wetsuit up into the air.
2. A man in a wetsuit is throwing a toddler up in the air and is ready to catch him.
3. A man in water throwing a little boy up in the air and waiting for him to come down so he can catch him.
4. The man is in the pool and throwing a small boy into the air.
5. While water droplets fly, a man throws a little boy up in the air.

Generated Caption:

Man throws little boy up in the air.

Actual Caption:

A woman and a baby waiting with luggage
A woman monitoring a baby as it plays with nearby luggage .
Near the seats , a woman has spread out green blanket to sit on floor with baby .
Woman and baby sitting on floor in waiting room .
Woman and baby waiting at an airport .

Generated Caption:

woman has spread out green blanket to sit on floor with baby

Actual Caption:
1. A dog is running across a desert with bushes around him.
2. A dog running across the sand
3. A dog running up a sandy hill
4. A little brown dog is running fast on a sand dune during the day
5. Small, hairy dog running in the sand.

Generated Caption:
Hairy dog running in the sand.

Actual Caption:
1. A basketball player performing a lay-up.
2. A boy in a blue basketball uniform, number 13 and a boy in a white basketball uniform, number 23 jump for the ball
3. A man in a white uniform jumps while holding a basketball as another in blue blocks him
4. Basketball player wearing a white, number 23 jersey jumps up with the ball while guarded by number 13 on the opposite team
5. The man in white is playing basketball against the man in blue

Generated Caption:
Number 23 jersey jumps up with the ball while guarded by number 13 on the opposite team

Fig 4. Images With Generated Caption

Although the results are promising, some limitations were identified. The model occasionally struggles with rare objects or ambiguous scenes, leading to captions that are partially accurate but lack detail. Infrequent repetition of phrases or generic descriptions points to the need for broader vocabulary coverage during training. These

issues could potentially be addressed by augmenting the dataset or incorporating attention mechanisms to refine contextual understanding.

5. CONCLUSION AND FUTURE WORK

In conclusion, this study presents a unprecedented approach to image captioning by integrating Vision Transformers (ViT) for feature extraction with Long Short-Term Memory (LSTM) networks for caption generation. The ViT effectively enhances the model's ability to capture complex visual relationships, while the LSTM decoder translates these visual features into coherent and contextually accurate captions. Applied to the Flickr8k dataset, the proposed method demonstrates strong performance across multiple BLEU score metrics, achieving near-perfect results, particularly at the unigram level (BLEU-1: 0.9961). The model's high scores reinforces its ability in generating captions that nearly align with human-annotated references, showcasing both accuracy in word selection and grammatical coherence. While a slight decline in performance is observed for longer n-grams, as reflected in the BLEU-4 score of 0.9901, the model consistently produces fluent and meaningful sentences, marking a significant improvement over traditional CNN-RNN architectures in capturing intricate visual details and generating precise image descriptions.

Future work can focus on broadening the scope and improving the model further. Model can be testes on bigger and diverse models, such as MS COCO or PASCAL VOC, will help validate its generalizability and applicability to more complex scenarios. Methods to enhance the diversity and creativity of captions, such as beam search with diversity penalties or integrating generative adversarial networks (GANs), could be explored to create more engaging outputs. Complementing BLEU scores with additional evaluation metrics like METEOR, CIDEr, and ROUGE will provide a more comprehensive analysis of caption quality. Furthermore, incorporating attention-based mechanisms and leveraging pretrained language models like GPT or BERT for caption generation could improve both semantic richness and the ability to focus on relevant image components. Deploying the model in real-time applications, such as tools or applications to help the visually impaired., is another promising direction. Detailed error analyses to address challenges like captioning rare objects or ambiguous scenes, combined with multimodal learning approaches that integrate audio or textual metadata, can significantly enhance the robustness and versatility of the system. By addressing these areas, the proposed methodology can be further refined, establishing a stronger foundation for advancements in image captioning.

References

6. Cheng, L., Wei, X., Mao, X., Liu, Y., Miao, C.: Stacked visual-semantic attention for image caption generation. IEEE Transactions on Image Processing 29, 3576–3590 (2020).
7. Amirian, S., Rasheed, K., Taha, T.R., Arabnia, H.R.: Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. Neural Processing Letters 52, 123–139 (2020).
8. Ding, G., Chen, M., Zhao, S., Chen, H., Han, J., Liu, Q.: Neural image caption generation with weighted training and reference. IEEE Transactions on Neural Networks and Learning Systems 32(12), 4875–4886 (2019).

9. Li, B., Munoz, J.P., Rong, X., Chen, Q., Xiao, J., Tian, Y., Arditi, A., Yousuf, M.: Vision-based mobile indoor assistive navigation aid for blind people. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(3), 742–755 (2018).
10. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Lessons learned from the 2015 MSCOCO image captioning challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 123–132. IEEE, Boston (2017).
11. Tiwary, T., Mahapatra, R.P.: An accurate generation of image captions for blind people using extended convolutional atom neural network. Neural Networks 122, 198–210 (2022).
12. Wu, C., Yuan, S., Cao, H., Wei, Y., Wang, L.: Hierarchical attention-based fusion for image caption with multi-grained rewards. IEEE Access 8, 44265–44273 (2020).
13. Ma, J., Jiang, X., Jiang, A., Yan, J., Jiang, J.: Image matching from handcrafted to deep features: A survey. Pattern Recognition 109, 107601 (2021).
14. Wang, E.K., Zhang, X., Wang, F., Wu, T.Y., Chen, C.M.: Multilayer dense attention model for image caption. Journal of Artificial Intelligence Research 34(4), 76–89 (2019).
15. Karim, F., Majumdar, S., Darabi, H., Chen, S.: LSTM fully convolutional networks for time series classification. Neural Processing Letters 19(4), 567–578 (2018).
16. Zhao, Y., Takaki, S., Luong, H.T., Yamagishi, J., Saito, D., Minematsu, N.: Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder. IEEE Transactions on Speech and Audio Processing 29(1), 1123–1133 (2020).
17. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. Journal of Artificial Intelligence Research 33(2), 411–420 (2020).
18. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models, and evaluation metrics. Journal of Artificial Intelligence Research 47, 853–899 (2013).
19. Mouineeshwari, R., Pethalakshmi, A.: Automatic image caption generation using deep learning. In: Proceedings of Springer International Conference, Lecture Notes in Computer Science, vol. 11155, pp. 123–135. Springer, Heidelberg (2019).
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Müller, D.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
21. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 315–323. IEEE, Boston (2015).
22. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 312–322. IEEE, Boston (2015).
23. Papineni, K., Roux, L., Zhou, L.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318. ACL, Stroudsburg (2002).
24. Chen, S., Cherry, C.: A systematic comparison of smoothing techniques for sentence-level BLEU. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Short Papers, pp. 362–367. ACL, Stroudsburg (2014).
25. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning phrase representations via tuning language models. In: Proceedings of the 43rd International ACM SIGIR Conference on

Research and Development in Information Retrieval, pp. 1265–1274. ACM, New York (2020).