ICIC Express Letters
Part B: Applications
Volume **16**, Number **7**, July **2025**

ICIC International ©2025 ISSN 2185-2766
pp. 735–742

ENHANCING IMAGE CAPTIONING THROUGH SEMANTIC EMBEDDING AND ATTENTION MECHANISM

Dessy Santi[1,3], Amil Ahmad Ilham[2,*], Syafaruddin[1] and Ingrid Nurtanio[2]

[1]Department of Electrical Engineering
[2]Department of Informatics
Universitas Hasanuddin
Jalan Poros Malino Km 6, Bontomarannu, Gowa, Sulawesi Selatan 92171, Indonesia
santid21d@student.unhas.ac.id; {syafaruddin; ingrid}@unhas.ac.id
*Corresponding author: amil@unhas.ac.id

[3]Department of Information Technology
Universitas Tadulako
Jalan Soekarno Hatta Km 9, Palu, Sulawesi Tengah 94148, Indonesia
dessy@untad.ac.id

Received September 2024; accepted December 2024

Abstract. *In fact, attention mechanisms and word embedding techniques are the advanced technologies provided by this generation of captioning. The goal of this study is to create an image captioning model that utilizes FastText with attention mechanisms and semantic embeddings to help generate accurate and contextually appropriate descriptions. Attention mechanisms are important for the model since they concentrate on important parts of images just like human beings do. Subword features in FastText make it possible to handle unknown words thus making the model more flexible and adaptable to new information. Experimental results on Flickr30k dataset indicate that ResNet-152 with Subword Mechanism, LSTM (Long Short-Term Memory) and Attention has achieved better performance as compared to other models, scoring a BLEU 1 score of 0.58 and METEOR score of 0.23. The ResNet-152 configuration works best among others hence providing a solution for image captioning tasks that can be used in production systems.*
Keywords: Attention mechanism, FastText, Image captioning, LSTM, Semantic embedding

1. **Introduction.** The generation of image captions enables computers to produce descriptions of images. A significant advancement in this field is the use of attention mechanisms, which allow models to focus on important parts of an image, similar to human attention to detail. Without attention, models produce general descriptions and miss specific details. The attention area method [1] associates adverbs with image regions. Research directs the model to relevant parts of the image to generate words. [2] combines Bahdanau's attention mechanism with two trained convolutional neural networks to predict specific image captions. [3] demonstrates the effectiveness of the attention approach with the Show, Attend, and Tell model. Despite these advancements, predicting contextual data remains a challenge due to the complexity of language and handling words that are not in the training dataset. This research proposes an image captioning model that uses attention mechanisms and semantic embeddings with FastText subword techniques to handle new words not present in the training dataset. By leveraging semantic relationships, the model is expected to generate more coherent and contextual texts. The academic contributions of this paper significantly advance the field of image captioning by integrating FastText subword embeddings with Bahdanau attention mechanisms. This approach effectively addresses the common issue of missing target words in the lexicon

DOI: 10.24507/icicelb.16.07.735

and enhances the model's ability to generate detailed and contextually relevant captions. By leveraging the strong visual features of ResNet-152 and the rich semantic information in FastText, the model produces more cohesive and descriptive phrases. The proposed approach is better compared to the current algorithms; this has been tested with extensive comparative experiments using BLEU and METEOR metrics, showing potential to enhance human-like comprehension in picture description tasks. Section 2 reviews the literature. Section 3 discusses the methodology, including preprocessing, semantic embedding, visual feature extraction, and LSTM with attention to the caption generation process. Section 4 presents the results and evaluation, and Section 5 examines the impact of semantic embedding and attention mechanisms on image captions.

2. **Literature Review.** Image captioning is a process for obtaining a textual description of an image with the use of computer vision and natural language processing. Attention mechanisms and semantic embeddings are recent neural network improvements instrumental in this task. This literature review explores key developments and methodologies in the field.

The attention mechanism is a key development that makes it better for computers to create accurate and relevant descriptions of images. The idea started with a project [4] called "Show, Attend, and Tell", where researchers figured out how to enable models to focus on important parts of an image when generating captions. Later on, new advancements made these models even more flexible [5]. For instance, a method that allows models to learn about new objects that they have not seen before was introduced, helping them improve their understanding [6,7]. Another improvement used a technique that identifies various objects in images more effectively, which helps in creating better descriptions [8]. Additionally, different models, like LSTM (a type of neural network), can focus on specific objects and how they relate to each other in the image [9-12]. The research by Liu and Deng in [13] offers superior image quality while preserving important details through Convolutional Block Attention Module (CBAM) with attention mechanism. All these methods combined strengthen the ability of these models to produce detailed and precise descriptions of what is happening in an image.

The text talks about ways to help computer programs understand and generate human language better by using something called semantic embeddings. These are essentially mathematical representations of words that capture their meanings in a form that computers can work with [14]. There are several methods to create these representations. One popular approach is called Word2Vec [15], which learns the relationships between words by analyzing large amounts of text. Another method, GloVe, looks at how often words appear together in texts to create its representations [16]. FastText is another technique that is particularly good at understanding rare words or words that may not even be in its training data [17]. The text also highlights how these word representations can improve specific applications, like generating descriptions for images [18]. For instance, one study combined FastText representations with a type of neural network (LSTM) to produce better image captions [19]. Other research explored combining visual features from images with these word representations to improve the overall model's performance [20,21].

Researchers have developed models that use two key ideas: attention mechanisms and semantic embeddings. The adaptive attention model works by looking at important parts of an image and also considering the words it is already used in its captions [22,23]. This helps it generate captions that are more relevant to what is in the image. Semantic embeddings add deeper context to the words the model generates, while the attention mechanism helps it zero on the relevant parts of the image for better captions [24,25].

There have been improvements in creating descriptions for images, but there are still some tough problems to solve. One issue is understanding complicated scenes that have a lot going on, like multiple objects interacting with each other. Another problem is making sure the text is varied and interesting, instead of just repeating the same phrases. It is also hard to useless common words since the training data might not include them. Looking ahead, researchers could try using more advanced methods to better focus on the details of what is happening in an image. They could also work on techniques that help produce a wider range of descriptions and include rare words to make the captions sound more natural and accurate.

3. **Methodology.** The methodology for developing an image text model using CNN with ResNet-152 architecture, FastText semantic embedding and LSTM with an attention mechanism is structured through several main steps. Below is a detailed description of the techniques used in Figure 1.

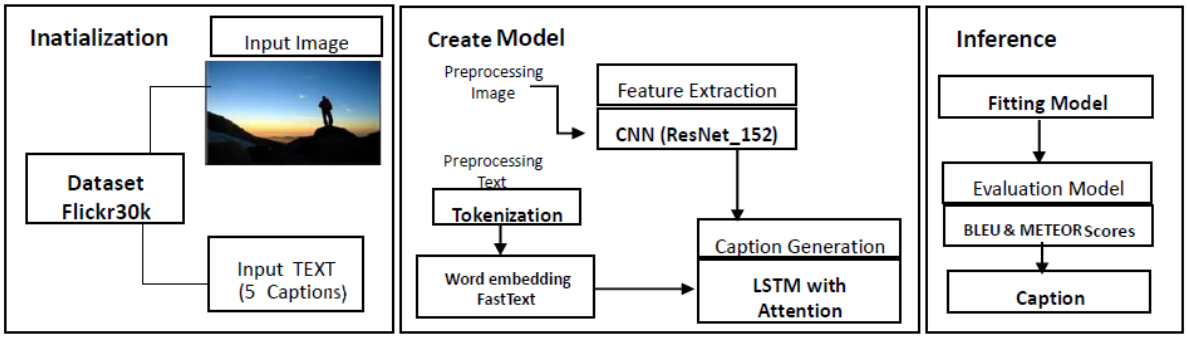


FIGURE 1. Overview of the workflow for the image captioning model

3.1. **Dataset.** The dataset used in this research is Flickr30k [26], which comprises 31,783 JPEG images of various sizes and shapes, each with five captions. The dataset is divided into three subsets: 25,783 images for training, 3,000 images for validation, and 3,000 images for testing. An example of an image and its captions from the Flickr30k dataset can be seen in Figure 2. The preprocessing stage involves removing punctuation, transforming all words to lowercase, and tokenizing the text to break it into individual words.

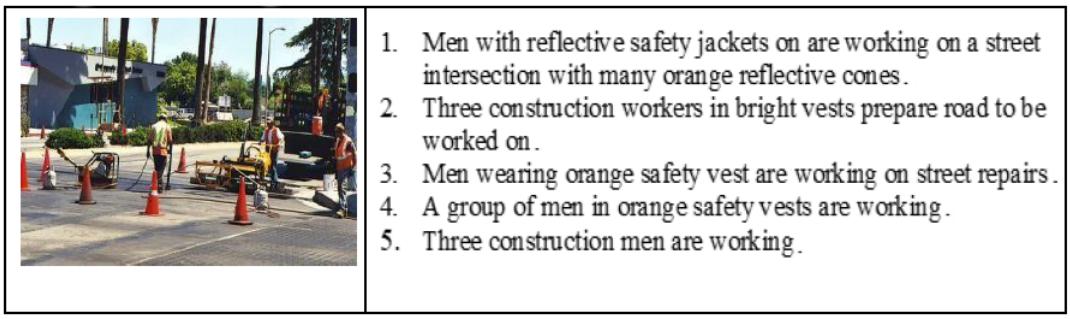


FIGURE 2. Image and caption samples from Flickr30k dataset [26]

3.2. **Semantic embedding.** The FastText model is a tool that helps understand the meanings of words by looking at their parts, which is useful when working with captions for pictures, like those in the Flickr30k dataset. This comes in handy when you encounter unusual or new words that the model has not seen before. The model uses a specific training method called "skip-gram", runs for a set number of 5 cycles, and operates at a learning speed of 0.1. It represents words using vectors that have 100 dimensions and

considers a small number of surrounding words (5 words before and after) to understand context. If a word is not directly recognized by the FastText model, it uses parts of the word (subword representations) to create a meaningful representation of that word.

1) **Subword Information.** For example, the word 'reading' is broken down into subword bi-grams: ["<r", "re", "ea", "ad", "di", "in", "ng", "g>"] ["<r", "re", "ea", "ad", "di", "in", "ng", "g>"] ["<r", "re", "ea", "ad", "di", "in", "ng", "g>"].
2) **Vector Representation.** Suppose each subword has a vector (v). The vector for the word 'reading' is Vreading = V $<r$ + Vre + Vea + Vad + Vdi + Vin + Vng + Vg >.
3) **Training Objective.** The objective function for training a word embedding model is given by Equation (1). The equation's first component maximizes the word's positive connection with pertinent contexts, while its second component reduces the word's negative association with irrelevant contexts (negative samples).

$$\sum_{context} \log \sigma\left(\vec{v}_{context} \cdot \vec{v}_{reading}\right) + \sum_{i=1}^{k} E_{v_{neg} \sim P_D(v)} \left[\log \sigma(-\vec{v}_{neg} \cdot \vec{v}_{reading})\right] \tag{1}$$

where $v_{context}$ and $v_{reading}$ are the embedding vectors for the context and the target word 'reading', respectively. The sigmoid function (σ) is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, $P_D(v)$ is defined as the word probability distribution within the corpus, v_{neg} is the embedding vector for the negative sample, chosen according to the word probability distribution $P_D(v)$ in the corpus and $E_{v_{neg} \sim P_D(v)}$ represents the average of the values calculated for several negative samples (k samples).

3.3. **Visual feature extraction.** A CNN with the ResNet-152 architecture transforms images into rich vector representations that capture the important visual elements of the images.

3.4. **LSTM with attention mechanism.** Bahdanau attention [24] is used to focus on specific parts of the visual features of the image when generating each word in the caption. Attention and text generation process:

1) **Initial State ($t = 0$).** The visual features from the CNN are used to generate the initial hidden state of the LSTM.
2) **First Word Generation.** The hidden state of the LSTM is used along with the visual features to generate the first word.
3) **Attention Mechanism.** Attention calculates the relevance score between the current hidden state and the visual features, generating a context vector that assists in selecting the next word.
4) **Generating the Next Word.** This process is repeated for each word in the caption, with the LSTM generating the next word based on the previous hidden state, the current word embedding vector, and the context vector from the attention mechanism.
5) **Generation Process.** Combining the extracted visual features with word embeddings from FastText and the context vector from the attention mechanism, the LSTM model generates words one by one to form a complete image caption.

3.5. **Evaluation model.** The evaluation metric used is BLEU (Bilingual Evaluation Understudy) [27] is based on n-gram precision and is calculated as the geometric mean of n-gram scores from BLEU 1 to BLEU 4. For each n-gram w_n in the candidate translation a, the frequency of occurrence $c_a(w_n)$ is compared to the maximum frequency of that n-gram among all references $\max_{j=1...|b|} c_{b_j}(w_n)$. The minimum value of these two frequencies is used to ensure that the occurrence of n-grams in the candidate is not counted more than the maximum occurrence in the references. Equation (2) measures

n-gram precision as the ratio between the sum of the minimum frequency values of matching n-grams to the total number of n-gram frequencies in the candidate.

$$BLEU_n(a,b) = \frac{\sum_{w_n \in a} \min(c_s(w_n), \max_{j=1...|b|} c_{b_j}(w_n))}{\sum_{w_n \in a} c_a(w_n)} \tag{2}$$

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [28] measures the quality of machine-generated image captions by comparing them with human reference texts. It combines precision (P) and recall (R) in the form of a harmonic mean (F-mean). A penalty factor reduces the F-mean score based on word order. Parameters γ and β control the magnitude of the penalty, and *chunks* and m represent the ratio between the number of chunks found in the translation and the total matching unigrams (m). The combination of both provides a METEOR score that reflects both content similarity and word order conformity, as shown in Equation (3).

$$\text{Score}_{\text{METEOR}} = \left(\frac{(1+\alpha)\cdot P \cdot R}{\alpha \cdot P + R}\right) \times \left(1 - \gamma \cdot \left(\frac{chunks}{m}\right)^{\beta}\right) \tag{3}$$

4. **Result and Discussion.** This architecture combines visual and textual features, utilizes semantic embeddings, and employs attention mechanisms to produce more accurate and meaningful image captions. The generated caption is 'a man reading a newspaper at an outdoor restaurant' as shown in Figure 3.

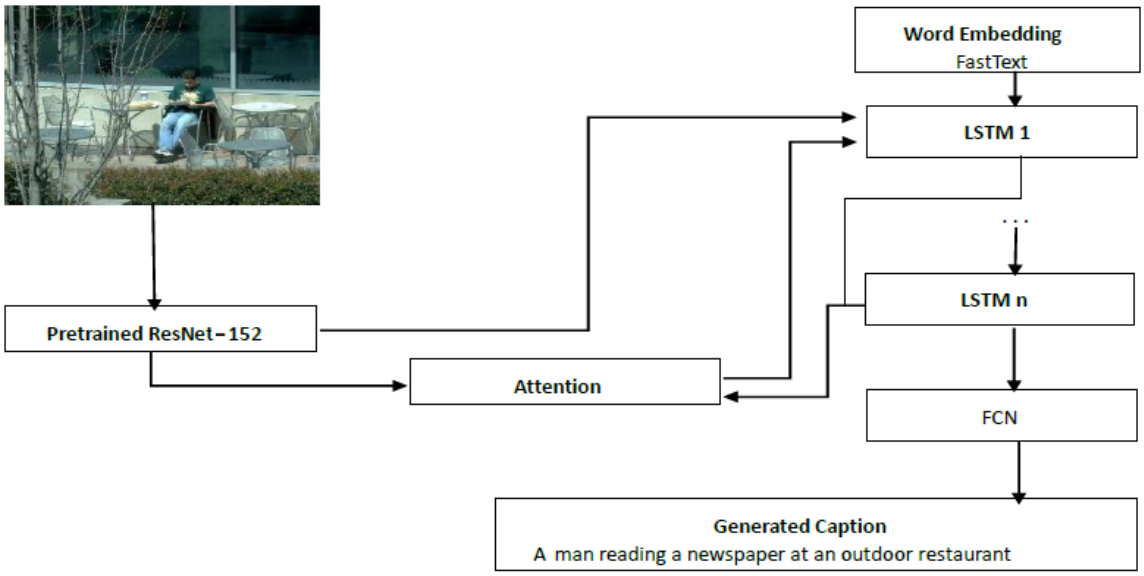


FIGURE 3. CNN and LSTM model with local attention mechanism

The testing dataset comprises 3,000 images, representing 10% of the total Flickr30k dataset. The BLEU and METEOR scores differ for each image. Table 1 illustrates three examples of generated captions compared to the ground truth, including the BLEU and METEOR scores for each image. The model can recognize images and rare words such as "grass", and accurately identify objects such as "trumpet" and "white shirt". Table 2 shows that the proposed model (ResNet-FastText-LSTM Attention) excels in BLEU 1 and BLEU 2. However, the Inception-resnetv2-Glove-RNN performs better in BLEU 3 and BLEU 4. For METEOR, the proposed model has the highest score, demonstrating its superiority in generating accurate and relevant descriptions, especially in the initial stages of text generation.

The quality of FastText embeddings also impacts performance if the embeddings do not align with the specific task or domain, the model will struggle to generate accurate

TABLE 1. Comparison of ground truth and generated caption with BLEU and METEOR scores

Image	Comparison of ground truth and generated caption	BLEU and METEOR scores
	**Generated Caption:** **A white dog is running on the grass.** **Ground Truth:** • A dog prepares to catch a thrown object in a field with nearby cars. • A white dog is ready to catch a yellow ball flying through the air. • A white dog is about to catch a yellow in its mouth. • A white dog is about to catch a yellow dog toy. • A white dog running after a yellow ball.	**BLEU 1** 0.5453 **BLEU 2** 0.5053 **BLEU 3** 0.4587 **BLEU 4** 0.4028 **METEOR** 0.1041
	**Generated Caption:** **A man in a black shirt is playing a trumpet.** **Ground Truth:** • A man in a black sweater with a red undershirt playing a trumpet in front of a lighted Christmas tree. • The man playing a trumpet in front of a Christmas tree looks intense. • A man in a black sweater and red shirt plays the trumpet by a tree. • A guy with a beard wearing his glasses is blowing on a trumpet. • A musician playing some smooth music.	**BLEU 1** 0.66 **BLEU 2** 0.2928 **BLEU 3** 0.087 **BLEU 4** 0.048 **METEOR** 0.133333
	**Generated Caption:** **A man in a white shirt is cooking in a kitchen.** **Ground Truth:** • Two men one in a gray shirt, one in a black shirt, standing near a stove. • Two guy cooking and joking around with the camera. • Two men in a kitchen cooking food on a stove. • Two men are at the stove preparing food. • Two men are cooking a meal.	**BLEU 1** 0.5855 **BLEU 2** 0.2774 **BLEU 3** 0.1858 **BLEU 4** 0.0874 **METEOR** 0.157895

TABLE 2. BLEU and METEOR scores

Evaluation metric	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR
ResNet-FastText-LSTM Attention (Our Model)	0.58	0.49	0.23	0.15	0.23
Inception-resnetv2-Glove-RNN [24]	0.561	0.501	0.494	0.445	–
Glove-LSTM [25]	0.53	0.24	0.10	0.46	0.21

captions. Figure 4 shows a consistent decline in the loss value from the start to the end of the training period. The rapid early improvement indicates that the model adjusts its weights to minimize prediction errors. The steady decline suggests that the learning rate and other training parameters were appropriately set for the initial training phase.

5. **Conclusions.** This research developed an image captioning model utilizing attention techniques and semantic embeddings with FastText to improve the model's capability to produce accurate and contextually relevant descriptions that are as natural as human language. The attention mechanism has proven to improve the model's focus on important elements within images. The use of FastText, particularly its subword features, offers advantages in processing unknown words during training, thereby increasing the model's flexibility and adaptability when encountering various types of new data. Future research could focus on making the models handle messy or unclear images better, using more

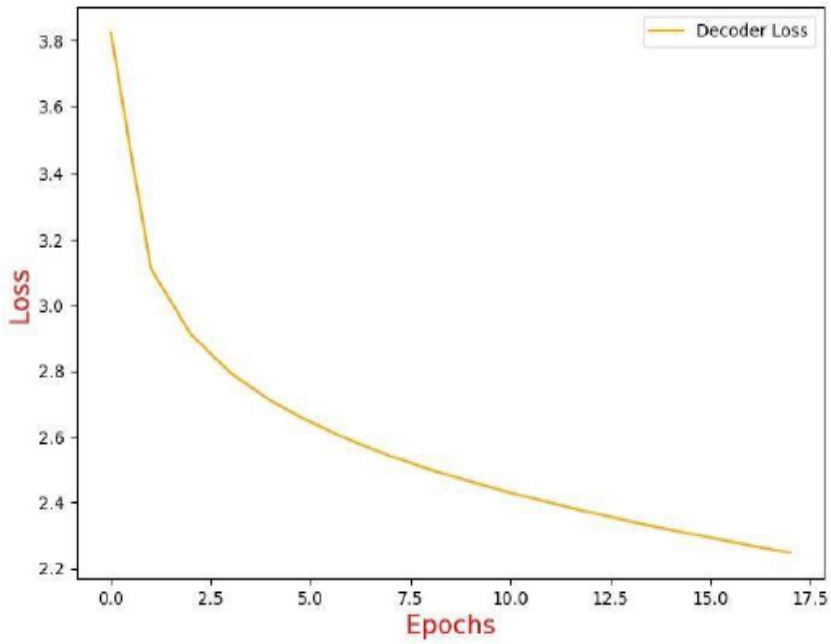


FIGURE 4. Decoder loss

advanced attention techniques, and seeing how they could work with augmented reality to provide real-time descriptions of what users are seeing.

Acknowledgment. The authors express their gratitude to everyone who contributed to the research process, making its successful completion possible. This research was funded by Hasanuddin University's Faculty of Engineering through the BPI 2024 Scholarship.

REFERENCES

[1] M. Pedersoli, T. Lucas, C. Schmid and J. Verbeek, Areas of attention for image captioning, *Proc. of IEEE Int. Conf. Comput. Vis.*, pp.1251-1259, DOI: 10.1109/ICCV.2017.140, 2017.

[2] S. Ayoub, Y. Gulzar, F. A. Reegu and S. Turaev, Generating image captions using Bahdanau attention mechanism and transfer learning, *Symmetry*, https://www.mdpi.com/2073-8994/14/12/2681, 2022.

[3] H. Liu and T. Brailsford, Reproducing "show, attend and tell: Neural image caption generation with visual attention", *J. Phys. Conf. Ser.*, vol.2589, no.1, DOI: 10.1088/1742-6596/2589/1/012012, 2023.

[4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, Show and tell: A neural image caption generator, *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp.3156-3164, DOI: 10.1109/CVPR.2015.7298935, 2015.

[5] J. Su, J. Tang, Z. Lu, X. Han and H. Zhang, A neural image captioning model with caption-to-images semantic constructor, *Neurocomputing*, vol.367, pp.144-151, DOI: 10.1016/J.NEUCOM.2019.08.012, 2019.

[6] T. Yao, Y. Pan, Y. Li and T. Mei, Incorporating copying mechanism in image captioning for learning novel objects, *Proc. of the 30th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR 2017)*, pp.5263-5271, DOI: 10.1109/CVPR.2017.559, 2017.

[7] P. Anderson et al., Bottom-up and top-down attention for image captioning and visual question answering, *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp.6077-6086, DOI: 10.1109/CVPR.2018.00636, 2018.

[8] V. A. Bharadi and S. S. Alegavi, Using Luong and Bahdanau attention mechanism on the long short-term memory networks: A COVID-19 impact prediction case study, *Smart Comput. Self-Adaptive Systems*, DOI: 10.1201/9781003156123-1, 2021.

[9] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja and P. Gupta, Deep image captioning using an ensemble of CNN and LSTM based deep neural networks, *J. Intell. Fuzzy Syst.*, vol.40, no.4, pp.5761-5769, DOI: 10.3233/JIFS-189415, 2021.

[10] Y. Feng, L. Lan, X. Zhang, C. Xu, Z. Wang and Z. Luo, AttResNet: Attention-based ResNet for image captioning, *International Conference on Advances in Computing and Artificial Intelligence*, DOI: 10.1145/3302425.3302464, 2018.

[11] N. Patwari and D. Naik, En-de-cap: An encoder decoder model for image captioning, *International Conference Computing Methodologies and Communication*, https://ieeexplore.ieee.org/abstract/document/9418414/, 2021.

[12] Y. Chu, X. Yue, L. Yu, M. Sergei and Z. Wang, Automatic image captioning based on ResNet50 and LSTM with soft attention, *Wireless Communications and Mobile Computing*, DOI: 10.1155/2020/8909458, 2020.

[13] J. Liu and Z. Deng, Low-light image enhancement using asymmetric convolutional neural networks, *International Journal of Innovative Computing, Information and Control*, vol.20, no.2, pp.479-496, 2024.

[14] V. Atliha and D. Šešok, Pretrained word embeddings for image captioning, *2021 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, https://ieeexplore.ieee.org/abstract/document/9431465/, 2021.

[15] A. Deoras, T. Mikolov, S. Kombrink and K. Church, Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model, *Speech Commun.*, https://www.sciencedirect.com/science/article/pii/S0167639312000969, 2013.

[16] J. Pennington, R. Socher and C. D. Manning, GloVe: Global vectors for word representation, *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532-1543, DOI: 10.3115/v1/d14-1162, 2014.

[17] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, Advances in pre-training distributed word representations, *arXiv Preprint*, arXiv: 1712.09405, 2017.

[18] F. Liu, X. Ren, Y. Liu, K. Lei and X. Sun, Exploring and distilling cross-modal information for image captioning, *arXiv Preprint*, arXiv: 2002.12585, 2020.

[19] Z. Chen, Y. Huang and L. Wang, Augmented visual-semantic embeddings for image and sentence matching, *International Conference on Information Photonics*, https://ieeexplore.ieee.org/abstract/document/8802975/, 2019.

[20] X. Zhang, S. He, X. Song, R. W. H. Lau, J. Jiao and Q. Ye, Image captioning via semantic element embedding, *Neurocomputing*, vol.395, pp.212-221, DOI: 10.1016/j.neucom.2018.02.112, 2020.

[21] X. Xiao, L. Wang, K. Ding, S. Xiang and C. Pan, Dense semantic embedding network for image captioning, *Pattern Recognit.*, vol.90, pp.285-296, DOI: 10.1016/j.patcog.2019.01.028, 2019.

[22] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, Image captioning using DenseNet network and adaptive attention, *Signal Processing Image Communication*, https://www.sciencedirect.com/science/article/pii/S092359652030059X, 2020.

[23] J. Lu, C. Xiong, D. PariNh et al., Knowing when to look: Adaptive attention via a visual sentinel for image captioning, *Computer Vision and Pattern Recognition*, 2017.

[24] Y. Bhatia, A. Bajpayee, D. Raghuvanshi and H. Mittal, Image captioning using Google's Inception-resnet-v2 and recurrent neural network, *2019 12th Int. Conf. Contemp. Comput. (IC3 2019)*, DOI: 10.1109/IC3.2019.8844921, 2019.

[25] A. Singh, T. D. Singh and S. Bandyopadhyay, An encoder-decoder based framework for Hindi image caption generation, *Multimed. Tools Appl.*, vol.80, nos.28-29, pp.35721-35740, DOI: 10.1007/s11042-021-11106-5, 2021.

[26] *Flickr8K*, https://www.kaggle.com/datasets/shadabhussain/flickr8k?resource=download, accessed in 2024.

[27] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of Annu. Meet. Assoc. Comput. Linguist.*, pp.311-318, 2002.

[28] A. Lavie and A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp.228-231, 2007.