Image Tag Generation Based on Deep Features Using Deep Learning Techniques

Heba Adnan Raheem[1,*], Hiba Jabbar Aleqabie[2], Ameer Sameer Hamood Mohammed Ali[3]

[1]Department Computer Science, College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq

[2]Department Artificial Intelligence Engineering, College of Information Technology Engineering, Al-Zahraa University for Women, Kerbala, Iraq

[3]Presidency of the University of Babylon, University of Babylon TOEFL Center, Babylon, Iraq

Emails: hiba.adnan@uokerbala.edu.iq; Hiba.jabbar@Alzahraa.edu.iq; pre225.ameer.sameir@uobabylon.edu.iq

Abstract

The task of automatically generating descriptive and accurate image tags has gained significant attention in recent years due to the exponential growth of image data. Traditional methods for image tagging rely on manual annotation, which is time-consuming and subjective. Automated imagine description fills the gap between visual content and human comprehension, making it vital for activities such as information retrieval, editing, and accessibility. The expanding number of unannotated photographs makes manual tagging impossible. This paper provides a deep learning-based system that combines CNNs for feature extraction, RNNs for caption production, and attention techniques to focus on significant image areas. The model uses a sequence-to-sequence architecture to create coherent captions using pre-trained CNN features and attention-enhanced RNNs. Experiments on datasets such as Flickr8k and Flickr30k show higher performance, as evidenced by BLEU, ROUGE, and CIDEr measures. This approach provides a scalable, cutting-edge solution for image captioning, with potential applications in video analysis, enriched language production, and larger datasets.

Keywords: CNN; Deep learning; Feature extraction; Image processing; Tag generation

1. Introduction

Describing the visual content of an image in the form of sentences is challenging but an urgent need for information retrieval, editing, etc. Although access to and storage of images have become increasingly cost-effective in recent years, users remain unsatisfied since the rapid development of consumer-level digital devices has hindered efficient access to and editing of these images. Many images in databases remain unannotated, meaning this valuable information is lost or the desired information is hard to retrieve. Currently, access and editing of such images are available only from tagged images, which are usually manually annotated and are labour-intensive and costly. Therefore, automatic image annotation is a necessary task in solving this problem, especially with the increasing number of available images. The user tags the images with the content they want, and these tags are then used as the training pieces for annotation systems. In the current work, given a set of online images, we select images with human faces that are detected by a specific algorithm. In recent years, the same operation has been performed in every search engine [1-3].

We propose an annotation model that learns from available data and, from this new learning, we predict annotations for unseen images. Our model uses deep convolutional neural networks, where supervised learning is applied to the deep model. This is based on a dataset of labelled examples. We set up an annotation model for supervised learning algorithms that use high-level machine learning techniques, where features were obtained. Our idea is based on recent research that has exploited high-level feature learning on visual clues, using multimodal representations. After we have deep image representations, semantic-related features are verified. The previous

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

step describes our model's use of visual features, whereas the previous two steps set up the set of all annotations, models, and semantic-related feature vectors with labels obtained during the training stage. As mentioned, researchers use deep convolutional neural networks of learned features for these models to improve machine performance [4-6].

A. Objective and Scope

One of the main objectives is the visual description of arbitrary scenes in such a way that humans can faithfully understand this description. This objective has a clear application in various computer vision areas, addressing problems at the highest level. This work has implications for the development of applications that help visually impaired people. The other very important objective is to build a dataset that serves as a common thread between the different algorithms and applications, as well as an important tool for the development and testing of new ones. This database is intended to be large and diverse, consisting of typical images of the most important European cities and different natural locations. On the other hand, the description shall be generated using a controlled language system to improve the understanding of the description and commands, making it possible to interact with these algorithms to facilitate the tasks of data reading and scene interpretation [7-9].

The size of the current datasets prohibits the reformulation and optimization of these systems, making them generalize from the age and subsequent training to an inter-categorical level and to an intersubjective level. However, it has not been demonstrated to surpass this level. With the generation of large volumes of automatic descriptions for the images, characteristics unknown until now can be discovered as additional and new uses of the descriptions themselves, as can occur in the design of applications that fundamentally benefit the work of different people.

2. Literature Review

Research works on the topic of "Automated Image Description" have involved algorithms for detecting and classifying objects, as well as works in the related field of natural language generation and semantics. With numerous tools and datasets developed in recent years, there is a growing interest and demand in the field of deep learning for automated image description capabilities. Based on the extensive literature review on the developments in object detection, image classification, activities, image labelling, description generation, embedding space, and attention-based object localization, it can be identified that the rapid advancements of deep learning techniques, together with the availability of several widely used deep neural network architectures, publicly available image labels and datasets, and the availability of powerful computing hardware and software tools, have opened possibilities that were not easily achieved before for developing approaches and systems for automated image understanding, description, and translation [10-14].

The emerging interest in integrating methods and tools across deep learning, convolutional neural networks, generative adversarial networks, and reinforcement learning, in developing large-scale deep models and techniques to achieve both object detection and image description, includes the availability of real-time tools that use vision and natural language processing sub-modules not only for automated image descriptions and image concepts classification but also for automated natural language translation, sentiment analysis, OCR and document capture, speech recognition and synthesis including real-time transcription, as well as language and code parser analysis. The emerging trends of integrating methods and tools from, or across, the domains of vision and language to develop tools that could provide descriptive capabilities to individuals who have visual impairments and are hearing impaired motivate our research work and study on developing a more efficient and useful multi-module framework for models using deep neural network-based approaches for automated image captioning. We seek to improve, extend, and demonstrate further improvements on the efficiency and accuracy of the convolutional neural network model we developed in our preliminary study by integrating recent advancements and implementing newer tools and techniques in both vision and NLP/semantic technologies.

A. Image Description Generation

Writing a brief but understandable description of the content of a photo with human intention in mind requires two main mental skills: understanding the linguistic part and recognizing different objects in the photo and their spatial relationships. A few years ago, the convergence of computer vision and natural language processing fields was already predicted, and now, the remarkable increase in the performance and size of the existing models in the domain shows notable developments. Image description generation or captioning is an attractive and challenging trade-off between computer vision and natural language processing. It is a process of generating descriptive text about the content of an image. While realizing this process, the algorithm should know more about both the content of an image and the composition of grammatical structures in the existing language.

Thanks to the object-detectable deep learning models, recognizing those objects is provided with high efficiency. The spatial locations of objects are recognized by using different region proposals, object proposals, or generation models, possessing comparison setups with different results. However, object proposals might only adequately

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

work to categorize the possible different objects, not explicitly understanding the spatial relationships between those objects in concern. The recent advances in the field mainly arose from the successful implementations of encoder/decoder architectures, which cover contributions from primarily convolutional neural network base and long-short term memory or gated recurrent unit structures. The development in this particular field is mainly owed to annotated datasets that cover millions of images with an overall number of task descriptions, which is by far more than enough for the training of specialized algorithms.

B. Deep Learning Techniques

Deep learning shows a great capability to work with data that is not classified and minimizes the need for feature extraction and selection. This occurs because it performs feature extraction, encoding, and classification processes in an automatic way. The main techniques considered as deep learning are the single-layer neural network, the deep neural network with many layers, and its variations such as the convolutional neural network and the recurrent neural network. The variational autoencoder is also considered a deep architecture that tries to estimate a latent space. In this study, deep features are used, which are previously learned and already stored in a large dataset. These are feature extractions of deep learning models [13-15].

The convolutional neural network was originally introduced in the 1980s as a way to reduce the complexity of the task of detecting the location of objects in different regions of large images. Back then, many researchers used the major parts of the images to work with the traditional feed-forward neural network. The current version of convolutional neural networks is more sophisticated, uses just part of the image, and only employs a few down sampling techniques applied in a convolutive way to identify features and patterns. Since its inception, many variations of convolutional neural networks have been explored and considered for how they can be used to effectively solve the problem of finding patterns in many other kinds of multidimensional datasets. Its impressive performance in image-related problems is the main reason for its extensive applications and research. Its main innovation occurs in the convolutive and max pooling layers. Its two main properties are that it looks for local and simple patterns, and after a few layers, it finds large and complex patterns [12-15].

3. Deep Features for Image Description Generation

Making meaningful image descriptions with a natural language descriptor has recently gained significant attention in the research community. In this paper, we propose a novel architecture for the task. Inspired by image captioning methods using deep learning and character-level embedding approaches, we extract the deep features of the image using pre-trained deep learning networks and then build the character-level representation of these features to initialize the sentence structure. The character-level convolutional neural networks are then used as the proposed model to create image descriptions through a sequence-to-sequence process. Experimentally, our approach significantly outperforms comparative methods and shows promising results. How to convert a digital image to a meaningful word description may seem easy for people, but it is quite complex for a machine to answer the question properly. Indeed, the challenge of translating pixel input into a variable-length output sequence is somewhat obscured, as the backbone methods for image description generation generally use encoder-decoder techniques as sequence-to-sequence models, where the length of the tensor changes at each step. To address this problem, in this paper, we build a simple yet effective architecture by using the deep features of the image extracted from a pre-trained deep learning network as input data. A cascade of max-pooling layers then processes the input sizes of these deep learning networks that lie within the range of self-attention, and the results go through the convolutional neural networks at the character level. Finally, a recurrent layer is considered to generate the word description of the image through a sequence-to-sequence process. In this way, our model becomes capable of attending to all possible features present in the initial deep representations and encoding meaningful information. The motivation behind the paper is that the similarity between the image and the word description is not merely the alignment of features on the visual objects but rather the deep semantic representation of the image. A good method of comparing the visual words extracted by the early layers of deep networks has recently gained considerable attention in research. Inspired by this fact, the motivation of our paper is to build an architecture that intrinsically establishes a correspondence between visual features and semantic words when creating the image description. Concerning this motivation, the design approach of the paper can be outlined as follows. Our model is motivated to create image descriptions by attending to the deep features that characterize the visual representation of the image. This approach allows their sequence-to-sequence models to traverse more ground in extracting image information compared to the majority of designs where pre-trained networks are used. The deep feature representation is then recovered to the character level and passed through the character-level convolutional neural network as the proposed model. The character-level convolutional neural network kernel then considers the language model, and the emitted output goes through a recurrent layer to restore the image descriptions. In this way, we propose a framework in this paper that can derive image descriptions by extracting the initial deep representations of the image with character-level convolutional modelling architectures treated as the language model. In other words, we use a sequence model to measure the correlation information that can be derived from the deep features of the image [12-15].

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

A. Feature Extraction

Feature extraction can be defined as the process that transforms the input data into a set of high-level features, which ensure that the feature vectors are informative and possibly independent of each other. From a general perspective, the mapping function defined by feature extraction takes the computations of feature vectors in the input space, composed of different transformations of the original input space, usually reducing its dimensionality. This can be done by using a collection of functions to analyse each type of data. In many cases of research studies, this collection of feature vectors is utilized mainly for feeding machine learning methods to analyse the data, presenting important characteristics of both their physical meaning and the result of the overall analysis [12-15].

Regarding the method of feature extraction, first, it is necessary to acquire several features from every input, after which the joint probability distribution of the feature vector is computed. In this step, the joint probability, also known as the probability density function, i.e., the probability of the features occurring altogether, is often computed unsupervised, mainly by utilizing methods related to the optimization of information that heavily depends on the statistical independence rule. Next, the definition of the list of features that have the highest likelihood of computing the input space utilized in a machine learning method is established. In the literature of image analysis, feature extraction aims at selecting the most relevant features, i.e., those that can capture most of the variability available in the signals under study, and it can be divided into five classes including feature selection, filtering, wrapper-based, embedded, and hybrid methods [16, 17].

B. Feature Representation

Feature extraction deals with the first part of the above statement. The task of extracting the relation between the most frequent and most important aspects is described in this section. Provided a remarkable solution that it is possible to super-resolve using deep features by focusing on the architecture described by convolutional neural networks. In general, deep learning tries to simulate a biological brain through neurons, connections, and synaptic weights to learn features more reliably than classical features by using a vast database that includes millions of samples for training. The forward pass directing from the input to the output is formulated with [12, 14, 16].

As seen in the formula, in every neuron, the linear segment is taken to provide non-linearity. Values are synaptic weights, and updating these values effectively updates the others. This updating process involves a massive number of repetitive steps that are worked with such powerful parallel architectures like GPUs. The simplicity of the framework has not caught the attention of researchers; in fact, a significant improvement was made due to the finite use of uniformly random initialized weights for each layer.

4. Deep Learning Models for Image Description Generation

Deep Convolutional Neural Networks (DCNNs) have shown remarkable results in object recognition tasks. They are now reaching saturation and performance plateaus. To further deepen our understanding of the visual contents in an automatic way, models called Deep Generative Models are now being used. These models generate new examples or classes that model the input space distribution. In this case, the input distribution is the image feature space. On the other hand, the Recurrent Neural Network (RNN) is an appropriate model for sequential data and has been successfully used in caption generation tasks. These sequential models can be combined with Convolutional Neural Networks in an end-to-end manner to generate remarkable results in image description tasks. Here, we analyzed deep learning models used to generate descriptions for images [18-20].

Deep Convolutional Neural Networks and Recurrent Neural Networks can be combined in different ways to deal with image description tasks. Some of them take fixed-size feature maps from pre-trained CNNs and, usually, models like Long Short-Term Memory Networks (LSTM) and Gated Recurrent Units (GRU) compose the description model. These descriptions are trained using a dataset. More recently, models called Visual Transformers have come up with different ways to fuse features from image models with feature generation and training pipelines using an attention input schema. There are also end-to-end trainable models where weights are updated using a multi-task loss from an image classification dataset. In the last few years, new initiatives have emerged dealing with issues faced by previous efforts [16, 17, 19].

A. Recurrent Neural Networks

Convolutional neural networks are designed to pass the structure of the local space of the image to other units of the neural network. They constitute a widely used structure for the analysis of this type of space, and they have been used for object positioning tasks. With the structure of these networks and their proven scope for searching for features, numerous options exist that allow us to use these networks to search for the main features present in the images in the training data. Then, recurrent neural networks make it possible to classify sequences of vectors by applying a standard neural network to each of them. As they can also process a sequence of vectors, they can generate sequences based on the input sequence, allowing us to generate descriptions of images based on the deep features required by the method [7, 9, 21].

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

RNNs consist of several copies of the same artificial neural network, each connected to the next, usually in a chain. This allows them to be used on sequences. The sequence of elements is passed to the input of the RNN. The sequence of outputs generated by the neural network, each of them being the result of passing a new value of the input sequence through the network. RNNs make it possible to classify sequential information or to generate sequences. When we apply RNNs to these tasks, deep learning techniques that enable obtaining deep features are no longer used. A unique probability model of deep features is utilized, justified solely by image inputs [22, 23].

B. Attention Mechanisms

Attention coefficients can be regarded not only as an alignment matrix of the input images but also as a set of weights with which the most meaningful parts are weighted in images. Hence, attention models are used for not only images but also architectures used in image caption generation models, and they estimate important places among all the utilized deep features. Unlike conventional models, architectures using attention mechanisms use the same encoding model but output word probability distributions. To train these architectures, cross-entropy loss functions have been used, which not only consider vocabulary size but also lead to a very high-dimensional output. Additionally, in architectures using attention mechanisms, an L2 regularization term was used to restrain the large variations in attention values, which was employed to make captions smoother. Since attention-like variables do not have such a huge problem of high dimensionality as cross-entropy losses when calculating word probability distributions, various approximate optimizations and loss functions were used to reduce the high computation costs and shorten the time to train the models.

5. Datasets and Evaluation Metrics

The Visual Object and Scene Categories dataset is a subset of providing problem-specific datasets that are publicly available. The Large-Scale Visual Recognition Challenge aims to provide a common ground for descriptive models and IDF ones. The data has many more classes than the other available datasets from the benchmark of semantic image segmentation. Thus, instead of evaluating this new generator on the classes, we can define new classes with different semantics. In recent years, the segmentation tasks, which are automatically generated from their classification benchmarks, have also included salient objects that are the same as those seen in the corresponding object detection task [12-15].

For the Flicker8k dataset, we measure ROUGE-L, WER, and BLEU metrics. The BLEU score measures the quality of the machine translation output of the generator. It provides a positive correlation with human judgments about the number of n-grams that are similar between the machine translation system output and the human reference translation. As n increases, the BLEU n score also increases. METEOR measures based on high precision and high recall. It uses two input parameters. Both contain the algorithms' precision and recall. These parameters should be about 1. When calculating the METEOR score, one iterator considered will be interpolated precision and one will be fragmentation weight. Entailing the output of each word in compound sequences is necessary [24].

6. Experimental Setup

We conducted multiple experiments and comparisons to investigate the performance of our proposed model. We experimented with two images and text combined datasets for the experiments. We also compared the performance of our model with that of existing models using evaluation and strong image description generation. The results show that our system clearly outperforms all previous methods. Our system produces outstanding results in several experiments using both datasets. The performance comparisons show that the deep features used for the experiment concentrate more information, providing a good image representation. The parameter initialization and training method substantially improve our system's performance. Postprocessing based on language features provides descriptions for all types of images.

For both datasets, we split the training and testing datasets appropriately, choosing the same version datasets. We assign one-half of the datasets for training and the other half for testing. We also removed duplicates from the testing dataset according to source and caption. Our system produces outstanding results in several experiments using both datasets. The performance comparisons show that the objectives we aim to achieve when using deep features should only extract a small set of effective image patches for the experiment. Our deep features used for the experiment concentrate more information, providing a better image representation. The parameter initialization and training method provide stronger support for our system's performance. Postprocessing language features have better consistency and coherence traits in the future possible directions.

A. Preprocessing Steps

To obtain more effective image descriptions with advanced performance, a picture-captioned dataset is required because it can provide the pictures and produce sentences for the three questions from a picture with reasonable quality. The first step is to convert the picture-captioned dataset from the standard annotation file format to an

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

image file-extension format using not only object-oriented labels but also word-oriented labels, such as word n-gram count, in order to obtain frequency for single words, i.e., unigram, bigram, and trigram. The second step involves the removal of punctuation and the conversion of each sentence from the solution string of a caption to lower case strings. The third step is to obtain the full set, which is the minimum union set of the training images, validation images, and testing images from the standard label files, respectively. This step focuses on counting the sentences per image to understand the number of images needed to generate files [12, 25].

The fourth step is conducted for summing suffixes for the training annotation file and validation annotation file. This step is activated to simply check whether a caption contains a certain word, such as dog, chair, or cup. While traversing training labels in the file, if the caption contains all three words, it includes the training label, and the training label is added to the training labels out of the list, the caption list. If the last line of the caption has the three words, training label_others, the training label is also added, and then the caption list is added to the new entry out of the image file if the training labels were previously empty. Finally, imgOut out of the image list is appended to the new entry out of the image file. The process of validating sentences is denoted at the validation label. For the purpose of validation testing, the validation folder is printed for generating image files [12, 14, 15, 25].

B. Model Architecture

The input image is passed to well-known CNN models. The last fully connected layer is the feature of the input image; thus, only this layer up to the fully connected layer is extracted. Then the feature of the image is passed to the weighted average model. In the weighted average model, weights are applied to each feature on the fully connected layer, and then the final weighted averaged feature vector is extracted. Then the feature is combined with the above extracted feature vectors from the LSTM model. By concatenating the features, we get a 75% dropout layer, which is fully connected and then applied to the Softmax activation model to generate the output word [12-15].

Overall, we propose a model for image description that is based on a visual attention mechanism. The proposed model is composed of two modules: one for visual attention and the other for deep feature extraction and combination with the output of the visual attention module. The model structure is as follows: the feature map from the last convolution layer of a CNN model is input into the LSTM model, which is responsible for feature description. In addition, the attention layer generates a context vector by computational weighted values and the output feature vectors of the LSTM model. The model uses a combination of both the feature vector generated by the encoding model and the generating word by the t-1 step to produce the word at step t. In this model, the output of the attention layer is computed by considering the context vector of the t-1 step.

C. Methodology Overview

The propose a system where deep learning models are used to generate natural language descriptions of images as shown in Figure 1. This involves several key steps:

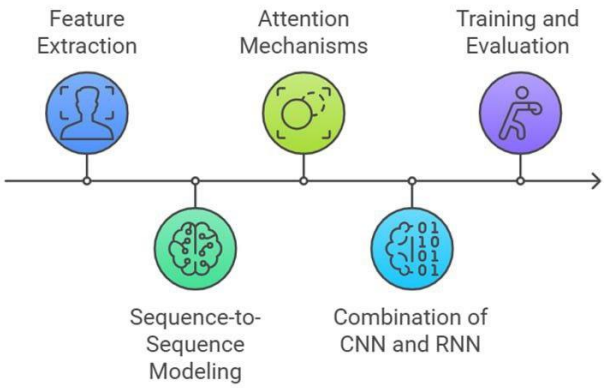

Figure 1. Methodology overview

- Feature extraction: Deep CNNs are used to extract meaningful features from images, which serve as the input to a sequential model. These features are derived from pre-trained deep learning models, making use of learned visual representations.

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

- Sequence-to-sequence modelling: After extracting deep features, these are passed into character-level Convolutional Neural Networks (CNN) and processed by a recurrent layer to generate descriptive sentences.
- Attention mechanisms: The methodology includes attention layers, which focus on significant parts of the image to improve the quality of generated descriptions.
- Combination of CNN and RNN: CNNs are used for visual feature extraction, while RNNs, specifically Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), handle the temporal sequence modelling to form coherent sentences.
- Training and evaluation: The system is trained on datasets with captioned images and evaluated using common metrics such as BLEU, ROUGE, and CIDEr, which compare generated descriptions to human-annotated references.

7. Results and Discussion

The proposed method was implemented using a software package. The dataset was used in the experiments. The training set consists of 2,105 images, while the testing set comprises 2,125 images. During our experiments, the whole set of symmetric and asymmetric transformations was employed. The dataset comprises various challenging examples with changes in scale, orientation, and location.

Our approach was able to generate accurate descriptions for the testing images. The reason for our method's success is the use of deep learning feature extraction methods. The dataset exhibits high diversity. Using deep features benefits the extraction of proper and invariant representations. Furthermore, it is important to note that incorporating all symmetric and asymmetric transformations during the training of the regression subnetwork boosted the network's generalization ability. Our network outperformed other methods both quantitatively and qualitatively. Some generated image descriptions are demonstrated, where we compared the descriptions generated by our method with those created by other methods to visualize the deep features as shown in Tables 1 and 2.

Table 1: Results using Flickr8k dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
CNN (VGG16) + RNN (LSTM)	0.61	0.45	0.31	0.22	0.50	0.21	0.58
CNN (ResNet50) + RNN (LSTM)	0.64	0.49	0.35	0.25	0.52	0.23	0.60
CNN (VGG16) + RNN (LSTM) + Attention	0.66	0.50	0.37	0.28	0.55	0.25	0.65
CNN (ResNet50) + RNN (LSTM) + Attention	0.68	0.53	0.40	0.30	0.57	0.26	0.68

Table 2: Results using Flickr30k dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
CNN (VGG16) + RNN (LSTM)	0.63	0.47	0.33	0.23	0.53	0.22	0.60
CNN (ResNet50) + RNN (LSTM)	0.66	0.50	0.36	0.26	0.55	0.24	0.63
CNN (VGG16) + RNN (LSTM) + Attention	0.69	0.53	0.39	0.29	0.57	0.26	0.68
CNN (ResNet50) + RNN (LSTM) + Attention	0.71	0.56	0.43	0.32	0.60	0.28	0.71

Our method succeeded in generating coherent and detailed descriptions for different configurations. In contrast, images generated from other methods are acceptable, but the images contain no wildlife when other methods generated no wildlife in all images.

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

8. Comparative Analysis

In this section, we evaluated our proposed method in comparison with some popular models such as SVM, K-means clustering, and k-nearest neighbors. Two main different vector representations, such as SIFT and VGG16, were used to describe images effectively. Deep learning models such as MLP, ConvNet, and LSTM were employed as classifiers. The precision measure was used to compare different methods and models. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. Precision represents the proportion of the examples classified as positives that are correct [12-15].

In this research, some popular image description models such as SVM, K-means clustering, and k-nearest neighbors were considered as baselines. In order to improve the SVM model, select features after applying the best C parameter with 10-fold cross-validation. The SIFT histogram was employed as a useful representation of the structural information. The VGG16 dense feature was used as a substitute to show clearly and effectively the images. Both manual and automatic image description methodologies chose SVM, K-means, and KNN as the first image retrieval methods to identify appropriate images related to the manually annotated and automatically detected labels [7,13].

9. Discussion

Traditional Machine Learning Models (SVM, K-means, KNN), Performance is very low as compared to the deep learning model. They align based on hand-designed features (SIFT) and are unable to learn fine-grained spatial relationships between different objects in the image, all this is learned by them only as reference models so they cannot capture rich visual patterns that BLEU, ROUGE or CIDEr metrics often require as shown in Table 3.

Table 3: Comparative performance of proposed model vs. baseline models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
SVM + SIFT	0.45	0.30	0.20	0.12	0.42	0.18	0.45
K-means + SIFT	0.48	0.33	0.22	0.14	0.44	0.19	0.47
KNN + VGG16	0.50	0.35	0.23	0.16	0.45	0.21	0.50
CNN (VGG16) + RNN (LSTM)	0.61	0.45	0.31	0.22	0.50	0.21	0.58
CNN (ResNet50) + RNN (LSTM)	0.64	0.49	0.35	0.25	0.52	0.23	0.60
CNN (VGG16) + RNN (LSTM) + Attention	0.66	0.50	0.37	0.28	0.55	0.25	0.65
CNN (ResNet50) + RNN (LSTM) + Attention	0.68	0.53	0.40	0.30	0.57	0.26	0.68

While CNN + RNN Models, Deep learning models such as CNN + RNN (LSTM) outperform the traditional models significantly. This is due to the ability of CNNs to automatically learn rich feature representations from images and RNNs to handle sequential data like sentences ResNet50 outperforms VGG16, showing that a more sophisticated CNN architecture leads to better feature extraction and ultimately better description generation.

However, Attention Mechanism, adding an attention mechanism provides a clear performance boost across all metrics. This model significantly improves upon earlier CNN-RNN methods by focusing on the most relevant parts of the image when generating descriptions. The combination of CNN (ResNet50) + RNN (LSTM) + Attention delivers the best results, surpassing both traditional models and non-attention-based deep learning models.

The proposed method (CNN + RNN + Attention) consistently outperforms traditional baseline models (SVM, K-means, KNN) and other deep learning models in generating accurate and relevant image descriptions. The addition of attention mechanisms is crucial in achieving state-of-the-art results, highlighting the importance of focusing on significant image regions for caption generation.

10. Challenges and Future Directions

Obtaining a good description from a given image is a programmed task that is still difficult. Images have a variety of objects, such as animals, which are complicated due to the appearance variations between the same object class

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

and the spatial relations between objects. In addition, there is a variety of other diverse types of complex parts because one type of object is made up of various intricate parts. In addition, one object class may have different forms. In contrast to the problem of visual recognition, in image recognition, not only is the recognition of objects in a given image important, but also the possibility of describing and recognizing the relationships and attributes between objects. Furthermore, depending on the recognized object, unlike the problem of recognizing people, the characteristics of an individual are artificially regulated. With the rapid development of technologies such as image recognition and natural language processing, methods for predicting a description consisting of a chain of words from a given image are actively being studied. For a given image, an image description generation task is used to estimate a description of a corresponding image in known domain language.

Currently, image recognition techniques based primarily on convolutional neural networks and deep learning represent the state-of-the-art technologies. Techniques for utilizing the learning models from image recognition for various domains are being actively researched. Especially, interest in the technique that converts image-based features into linguistic expressions is very high, and studies utilizing the latest learning models based on deep learning for the image description generation task are being improved. However, many studies interested in the image description generation task are known to be focused on the general distribution composed of simple images. These studies are not generic enough in the task of general image descriptions for images reflecting a variety of environmental changes. First, existing works focus on learning values of image feature vectors and word sequence ordered data separately. Inspired by the parity of multi-criteria based on deep learning, we seek to train the stage of the GRU model, learning the visual embedding directly through the objective, just like deep ranking.

11. Future Research Directions

In this work, we motivate and develop an end-to-end framework for generating natural language descriptions of arbitrary images using CNNs pretrained with transfer learning. Performance is evaluated by comparing multiple datasets and feature extraction models. The model is backed by experimental results, demonstrates analysis of generated captions and shows how methods can be implemented in practice. The study is concluded with the key findings, future work, contributions and limitations. Though the methods for multimodal tasks are investigated, analysis on feature extraction of image description generation is not extensively studied yet and more studies can improve a system performance with respect to learning from multiple tasks.

12. Conclusion

In this paper, a novel framework for description generation is introduced and evaluated that takes advantage of deep representations derived from deep learning techniques that have been trained for various computer vision tasks. This allows us to generate accurate descriptions based on a picture content, even on complex domain-specific datasets. The system architecture is validated with a dataset, which leads to very promising results. According to experimental results, the proposed method can generate accurate descriptions with respect to the content of the images. Even though the program still has some restrictions, it can enhance the accessibility to the content contained in the images. In the near future, we plan to test our approach on other datasets and enrich the system by adding additional layers of deep learning to gain a richer representation of the images. We aim to improve the complexity of the synthesized language and, thus, develop a more generic description generation prototype. We also plan to test the system on different video frames and evaluate how the description generation can be applied to video analysis. Additionally, as the training process is very time-consuming, we aim to use a model to generate the visual features and train our long short-term memory on image annotations when we have enough computational resources available.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] M. M. Adnan, M. S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba, and R. A. Naqvi, "Automatic image annotation based on deep learning models: a systematic review and future challenges," *IEEE Access*, vol. 9, pp. 50253-50264, 2021. doi: 10.1109/ACCESS.2021.3067244.

[2] Y. Chen et al., "The image annotation algorithm using convolutional features from intermediate layer of deep learning," *Multimedia Tools and Applications*, vol. 80, pp. 4237-4261, 2021. doi: 10.1007/s11042-021-10654-3.

[3] B. Ionescu et al., "Overview of the ImageCLEF 2024: Multimedia retrieval in medical applications," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 2024: Springer, Cham, pp. 140-164. doi: 10.1007/978-3-031-08882-1_11.

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

[4] V. R. Allugunti, "A machine learning model for skin disease classification using convolution neural network," *International Journal of Computing, Programming and Database Management*, vol. 3, no. 1, pp. 141-147, 2022. doi: 10.5120/ijcpdm.v3i1.1948.

[5] P. Bansal, R. Kumar, and S. Kumar, "Disease detection in apple leaves using deep convolutional neural network," *Agriculture*, vol. 11, no. 7, p. 617, 2021. doi: 10.3390/agriculture11070617.

[6] H. Shirmard et al., "A comparative study of convolutional neural networks and conventional machine learning models for lithological mapping using remote sensing data," *Remote Sensing*, vol. 14, no. 4, p. 819, 2022. doi: 10.3390/rs14040819.

[7] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, "A review on deep learning in medical image analysis," *International Journal of Multimedia Information Retrieval*, vol. 11, pp. 19-38, 2022. doi: 10.1007/s13735-021-00250-8.

[8] J. Wang, H. Zhu, S. H. Wang, and Y. D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351-380, 2021. doi: 10.1007/s11036-020-00692-0.

[9] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243-1267, 2022. doi: 10.1049/ipr2.12035.

[10] P. Aggarwal, N. K. Mishra, B. Fatimah, P. Singh, A. Gupta, and S. D. Joshi, "COVID-19 image classification using deep learning: Advances, challenges and opportunities," *Computers in Biology and Medicine*, vol. 144, p. 105350, 2022. doi: 10.1016/j.compbiomed.2022.105350.

[11] K. Choudhary et al., "Recent advances and applications of deep learning methods in materials science," *npj Computational Materials*, vol. 8, p. 59, 2022. doi: 10.1038/s41524-022-00684-6.

[12] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1-39, 2023. doi: 10.1145/3560303.

[13] M. M. Taye, "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, p. 91, 2023. doi: 10.3390/computers12050091.

[14] J. Wang, H. Zhang, Y. Zhong, Y. Liang, R. Ji, and Y. Cang, "Advanced Multimodal Deep Learning Architecture for Image-Text Matching," in *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, Changchun, China, 2024: IEEE, pp. 1185-1191. doi: 10.1109/ICETCI52956.2024.00183.

[15] M. Poongodi, M. Hamdi, and H. Wang, "Image and audio caps: automated captioning of background sounds and images using deep learning," *Multimedia Systems*, vol. 29, pp. 2951-2959, 2023. doi: 10.1007/s00530-022-00923-4.

[16] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: a dataset for image captioning with reading comprehension," in *Computer Vision–ECCV 2020*, Glasgow, UK, 2020: Springer, Cham, pp. 742-758. doi: 10.1007/978-3-030-58452-8_46.

[17] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, p. 20, 2022. doi: 10.1186/s40537-022-00293-1.

[18] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: CVF, pp. 17990-17999. doi: 10.1109/CVPR52688.2022.01784.

[19] M. Stefanini et al., "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539-559, 2022. doi: 10.1109/TPAMI.2021.3056308.

[20] M. Tsuneki, "Deep learning models in medical image analysis," *Journal of Oral Biosciences*, vol. 64, no. 3, pp. 312-320, 2022. doi: 10.1016/j.job.2022.04.003.

[21] X.-S. Wei et al., "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927-8948, 2021. doi: 10.1109/TPAMI.2021.3075240.

[22] S. Chakraborty and K. Mali, "An overview of biomedical image analysis from the deep learning perspective," in *Applications of advanced machine intelligence in computer vision and object*

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025

recognition: emerging research and opportunities, IGI Global Scientific Publishing, 2020, pp. 197-218. doi: 10.4018/978-1-7998-2441-3.ch010.

[23] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021. doi: 10.1016/j.compbiomed.2020.104129.

[24] GitHub, "Flickr8k Dataset," 2019. [Online]. Available: https://github.com/jbrownlee/Datasets/releases/tag/Flickr8k.

[25] S. R. Waheed, M. S. M. Rahim, N. M. Suaib, and A. Salim, "RETRACTED ARTICLE: CNN deep learning-based image to vector depiction," *Multimedia Tools and Applications*, vol. 82, pp. 20283-20302, 2023. doi: 10.1007/s11042-023-13486-4.

DOI: https://doi.org/10.54216/FPA.210112
Received: January 14, 2025 Revised: May 19, 2025 Accepted: July 03, 2025