Z10_Manuscript

by Rangaballav Pradhan

Submission date: 16-Jun-2025 06:11PM (UTC+0530)
Submission ID: 2700398863
File name: Z10_Manuscript.docx (1.55M)
Word count: 4144
Character count: 24483

IMAGE CAPTION GENERATOR USING CNN AND LSTM

Nigamananda Panda[1], Akashdeep Behera[2], Ganesh Das[3], Madhab Mohanty[4], Rangaballav Pradhan[5]

Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

[1]2141016090.z.nigamananda.panda2003@gmail.com
[2]2141002072.z.akashdeepbehera0106@gmail.com
[3]2141004078.z.ganeshchandradas663@gmail.com
[4]2141013301.z.madhavmohanty2003@gmail.com
[5]rangaballavpradhan@soa.ac.in

Abstract. As visual data becomes increasingly ubiquitous in online environments, there is a need for systems to automatically understand and describe the content of images. Annotating thousands of images takes a lot of time, and the annotations will likely be inconsistent, especially across very large sets of data. The proposed system describes an image captioning system that uses deep learning technology and a CNN-based image captioning method, which employs DenseNet201 to extract image features and LSTM to produce descriptive captions. The proposed system also includes a TTS engine to create speech output from the text, producing an enhanced user experience for aesthetics and accessibility for visually impaired users. To train and evaluate the proposed model, the Flickr8k images were used, which consist of RAW images and multiple human-annotated captions for each image. Our method's approach is based on generating caption accuracy, feature reuse, and generating speech output that sounds natural and easy for users to understand. The full model is coded in Python and implemented as a Jupyter Notebook in the Google Colab environment using TensorFlow. Experiments demonstrate that the system is capable of creating responses that are meaningful and grammatically possible as image descriptions. Though the current model works for moderate-scale datasets, there are many opportunities for improvement, including the use of attention mechanisms, multilingual systems, and real-time systems. Even though the developed systems seem to perform well as a reminder of human operators, the system's inaccuracies depend on the complexity of the image and the diversity of the captions, so continued improvement and testing will be necessary moving forward.

Keywords: Image Captioning, DenseNet201, LSTM, Text-to-Speech, Deep Learning, Flickr8k, Accessibility.

1 Introduction

With the rapid growth of visual content on social media sites, online databases, and personal storage methods, it has become more difficult to understand and save image information in a traditional manner. The inability to scale the description and classification of images has been detrimental to effective management of its content in a digital format. Traditional methods of annotation are inherently labor-intensive, slow, often error-prone, and practical for smaller datasets (with examples like thousands of annotated images from datasets like Flickr8k [1]). Recent breakthroughs in deep learning have allowed research into the automated description of images. While most currently available models use convolutional neural networks (CNNs), like VGG16 and ResNet50, common features of these architectures are: a lack of feature reuse, excessive CPU costs, and overfitting on moderate datasets. Because of its densely connected layers, DenseNet201 provides some advantages for feature propagation, provides a compact representation of image data, and accuracy improvements using smaller datasets that are adequate for scaffolding work like Flickr8k [7]. To produce descriptive captions, sequential information must be evaluated. Recurrent models, like Long Short-Term Memory (LSTM) networks, have shown strides in correctly modeling temporal dependence and ultimately producing coherent text [4]. This work brings together DenseNet201 as a source of rich image features and LSTM as a sequence generator of text. Additionally, a Text-to-Speech (TTS) module is added to voice the generated captions for the purpose of accessibility for users who are visually impaired [8]. The aim of this project is to show a practical image captioning system by building the model with Python and TensorFlow on Google Colab [2], [3]. Future extensions may include the use of attention mechanisms and multilingualism to enhance caption quality and target more users.

1.1 Motivation

With the exponential number of images that are produced and shared across all forms of social media, online platforms, computers, and personal devices in the digital age, the ability to automatically create semantic descriptions for images is becoming more relevant and necessary for organizations. Automatic image captioning (AIC) systems can serve a variety of purposes, including indexing content, search optimization, aiding digital libraries, and enhancing assistive technologies for visually impaired users.

While people can manually annotate and describe images, it is an operational burden that is subjective, time-consuming, and there are no definitive methods for managing the vast amount of data being generated. Automatic image captioning systems propose a practical solution, supported by advancements in deep learning. This paper has put forth a model that uses DenseNet201 for competition-grade and rich feature extraction of images, and LSTM networks for natural language caption generation of the extracted features. An automatic image captioning system could make it easier to manage and search for visual data in a more organized and searchable way. Additionally, limiting the burden on visually impaired users and giving them a narrated description of any image is a big step in enhancing digital accessibility [1], [2].

1.2 Objectives

This project aims to build an automated system to describe images through captions that are also converted into speech, which is more accessible, especially to the visually impaired. The method uses DenseNet201 to extract certain visual features beforehand, in combination with Long Short-Term-Memory (LSTM) Networks to create natural language captions. This also diminishes the need for human intervention for manual annotation and creating consistency across a large dataset. In particular, this project considers better accuracy and repeatability of captions that describe the image at a level other than just what is present in the image, but also the scene being depicted as well. The model can robustly produce captions regardless of changing factors in the images, such as content, lighting, or complexity. This project also considers minimizing the errors of captioning, polish/fluency, range of transpired syntax, and semantic descriptions. Additionally, as the work includes embedding text-to-speech capability, it takes it beyond basic image captions. These attributes of an automatic captioning system bring broader usability to assistive technology in creating a more interactive experience, not just a visual experience. Finally, the goal of this project will be to build a largely scalable, efficient, and realistic model for why it could have pragmatic application in healthcare, education, or utilization, digital media, content, and management technology, etc.

1.3 Original Contribution

This research offers a resource-efficient, easy-to-implement image caption system that brings together DenseNet201 for feature extraction, an LSTM network for language generation, and a Text-to-Speech (TTS) module for audio output. By using a small dataset and fewer resources, this study offers a platform for effective caption generation in a resource-constrained environment, contrary to the majority of past image captioning studies, which needed a large number of resources and a high-end platform, often hotel-level machine. In contrast to previous work, one of the more significant contributions is the use of the DenseNet201 deep convolutional neural network (D-CNN), which is a deep convolutional neural network that consists of many densely connected layers, improving the level of feature propagation and feature reuse. This enables better learning efficiency and reduces the redundancy of features in extraction processing when working with limited datasets. The features are passed to a Long Short-Term Memory (LSTM) model, which can take temporal dependencies of word sequences to generate accurate natural language captions. For enhanced accessibility for visually impaired users, it has a Text-to-Speech (TTS) component using gTTS for the generated captions, which generates audio of spoken outputs. The displayed output integrates all aspects of vision, language, and audio processing into one complete user experience. Additionally, the model is built and trained in Python and TensorFlow using the Google Colab platform, ensuring access to the system for researchers and developers without expensive computation resources. This study shows the real-world feasibility of using a deep learning-based image captioning system in low-resource settings and demonstrates the practical application of the model in education, accessibility, and assistive technologies.

1.4 Paper layout

The remaining of the paper is structured in a clear manner with headings that separate out the sections of the paper as follows. Section 2 provides a brief literature review. Section 3 discusses the proposed model. The experimental results are obtained and analysed in Section 4. Finally, Section 5 concludes the paper giving directions for further research.

2 Literature Survey

A significant area of growth in computer vision and natural language processing is known as image captioning, which aims to visually describe images in text. The initial developments in image captioning research were presented as a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units capable of receiving image features and producing

natural language captions as output. One relatively early representation of this architecture is referred to as "Show and Tell" (Vinyals, Toshev, Bengio, and Erhan, 2015) [1] wherein InceptionV3 was applied as a form of encoding/representation of the image as a CNN producing representations suitable as inputs, while the next stage was an LSTM soft-max decoding sequence. In contrast, "Show, Attend and Tell" (Xu, L., Laj, S., and Pirot, M.) [2] extended previous research with the attention element, allowing the model to analyze the image uncovered on-the-fly to identify pertinent areas to graphically attend/ concentrate on view the extraction of salient image regions which most relevant to the LSTMs task of generating natural language sequence text. In another notable dissertation, the principal observation in the Deep Visual-Semantic Alignment Model by Karpathy and Fei-Fei (2015) [4] illustrated an expected joint alignment between images and sentences, where bidirectional RNNs aligned image regions represented as features, aware of the several interacting sentences or sentence fragments at output levels against generated captions. This allowed the system/machine the ability to know how the various parts of that picture aligned to certain, and not so certain, parts of generated text; facilitating connectedness inherently or commonly understood as the principal component to bridge some of the gaps between vision and language semiotic methods. Moreover, Sutskever et al. [3] also contributed by increasing sequence-to-sequence models controlled using LSTM networks, thus explicitly maintaining coherence across very complex visual contexts with language generation. Johnson et al. developed the Dense Captioning model that uniquely captures region-based captioning by detecting multiple regions in an image, followed by individual captions for each region [6]. This system was trained on the Visual Genome dataset, allowing for greater understanding of the scene elements and the relationships between them, which opens opportunities for applications such as visual storytelling and question answering. Anderson et al. further enhanced captioning models developing the Bottom-Up and Top-Down Attention [7], using object-level attention based on Faster R-CNN for localization combined with a top-down LSTM caption generator function, allowing the model to focus on recognising objects with semantic representations rather than treating an image as a grid containing a series of pixels. In its entirety, the model yielded significantly greater accuracy for localising attention while dynamically generating captions. Recent advances have seen the introduction of several grounded architectures based on transformers in addition to preliminary pretraining of vision-language types of models. For instance, with the VirTex model recently developed by Desai and Johnson, groups visual features from image caption pairs without traditional classification pretraining [9]. This joint type of pretraining approach affords richer representations of visual information for both image captioning tasks and downstream tasks in vision and language. Just like Oscar (Object-Semantics Aligned Pre-training), as described by Li et al. [10], uses object tags for anchor points during vision-language pre-training to improve semantic knowledge and the performance on standard benchmarks, this paper applies the advantages of combining DenseNet201 [5], which is used for efficient visual features, DenseNet performs better in picture classification, with LSTM to generate sequences, leveraging the balance of performance and cost on computation. Using the Flickr8k dataset [2] provides a good balance between a finite, manageable collection of diverse images to be used for many epochs for training and evaluation. Easy accessibility using a Text-to-Speech (TTS) module based on gTTS, which uses the Google Text-To-Speech Engine [4] and deployment on Google Colab [10] simplifies experiments without requiring high-end hardware. These proactive advances showcase how the advancement of architecture and datasets are continuously evolving image captioning systems to be more accurate, faster, and pragmatically usable.

3 Proposed System

The system described in this paper aims to generate meaningful image captions and produce voice output in order to provide accessibility and general automation. The system utilizes the Deep Learning model DenseNet201, a comparatively deep and densely connected convolutional neural network (CNN) for extracting more detailed features from images. DenseNet201 promotes feature reuse and gradient flow, and so is able to be trained relatively easily even with limited computational power. Following the extraction of visual features with DenseNet201, those visual features are passed into a Long Short-Term Memory (LSTM) network, which takes the visual features in sequence and generates grammatically consistent and contextual captions one word at a time. To increase accessibility, especially for users who are visually impaired, a Text-to-Speech (TTS) module was incorporated using Google Text-to-Speech (gTTS), capable of transforming captions generated from the image into audible speech, allowing a user to interpret the image content through alternative means: audible speech. The whole model is trained using the Flickr8k dataset; it contains 8,000 images, each with five human-written captions, which provides a rich, well-balanced corpus of content suitable for training. We preprocess the dataset through resizing, normalization, tokenization, and sequence padding to ensure the data is compatible with the model architecture. The whole system is implemented using Python and TensorFlow and was trained using Google Colab, which is an online cloud-based development environment that does not require local hardware. The system is modular, scalable, and built on the basis of low resources. In this work, we have also demonstrated the effective combination of deep learning methods to develop an easy, quick, and lightweight solution for automatically generating image captions, with potential benefits to education, assistive technologies (for users with vision challenges), and multimedia resource management.

3.1 Methodologies

In this proposed image captioning system, we used the Flickr8k dataset [11], which contains 8,000 images with captions. The images are resized to fit in the Densenet201 and extracted for feature extraction. The captions are cleaned, tokenized, and padded. We embed the caption sequences and concatenate them with the image features, and put that through long short-term memory (LSTM) for descriptive text. Finally, the text-to-speech (TTS) module, using gTTS, converts the text into speech that allows for efficient visual representation, language generation, and voice output for improved accessibility.

Data-set Description: The Flickr8k dataset [11] used in this paper was obtained from Kaggle and contains a total of 8,000 images taken in the real-world setting that come with five human-annotated captions each, providing a total of 40,000 caption samples. Each image has a range of everyday scenes, making it appropriate for training image captioning models. Prior to ingestion into the model, the data underwent routine preprocessing steps, inclusive of image resizing, normalization, tokenization of the text, and padding of the caption sequences. The dataset was divided into training and validation data to properly learn and evaluate the performance. This ultimately enabled the use of DenseNet201 for feature extraction and LSTM for producing meaningful image captions.

3.2 Schematic layout of the proposed model

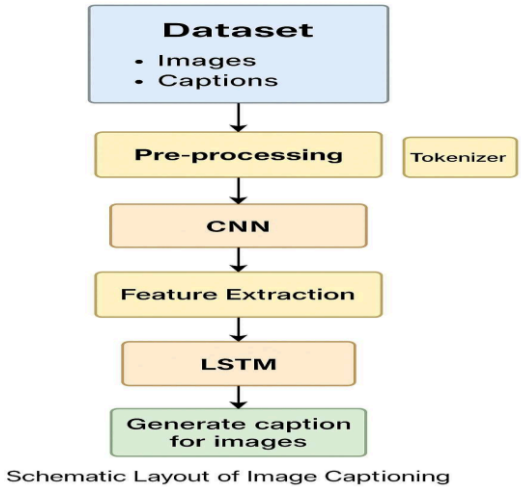

Fig. 1. Schematic layout

The diagram in Fig. 1 illustrates the different stages of deep learning image captioning in the steps of data pre-processing, splitting the data into development and test sets, extracting features using a CNN (Convolutional Neural Network) model, creating tokens using a tokenizer, and learning sequences using an LSTM (long short term memory) model. This descriptive flow diagram shows the connections between each stage and the process of automated generation of captions from image data.

3.3 System Requirements

The hardware requirements include a laptop with the appropriate operating system (like Windows), at least 8GB RAM (preferably 16GB or more), and ideally a GPU that is supported with CUDA, i.e., NVIDIA 1650 GTX. The software requirements include the programming language Python (version 3.7 or later). In addition to Python, you will need to install the following dependencies and libraries: Pandas, SK-learn, Matplotlib, NumPy, and Word Cloud. You will also need enough space to store the dataset, model, and run time files.

3.4 Proposed Models

The automatic image captioning model proposed in this study utilizes a deep learning architecture that combines both convolutional neural networks (CNN) and recurrent neural networks (RNN) techniques. Image features are first extracted by the CNN DenseNet201, which is well known for its structure of using all layers to connect to each other to allow higher feature reuse and avoid vanishing gradients. DenseNet201 can learn high-level visual semantics by the way it reconnects and places layers in a feed-forward structure, which essentially allows the model to represent visual information in a much denser or compact size. The feature vector is then passed on to an LSTM network to process the description in natural language. The LSTM process can take in features one at a time and model the long sequence of elements while taking in features to maintain context. It generates a sequence of words based on their predicted probabilities to provide grammatically relevant captions that correspond to the image. It uses a categorical cross-entropy loss function that measures the distance between the original word probabilities and the predicted probabilities (258). The model used an Adam optimizer to update model weights and biases. In order to make the captions more accessible, the text that was generated was converted to speech using the Google Text-to-Speech (gTTS) API for audio consumption. Finally, everything was run through Python and TensorFlow with the Google Colab platform, allowing for easy scaling and deployment. The final solution weighed a trade-off between accuracy and serving so as to serve well in education, assistive technology, and other cases.

4 Model Evaluation

The assessment of the image captioning model focuses on assessing its performance in learning and the quality of generated captions with standard training metrics such as categorical cross-entropy loss and word-level accuracy. Categorical cross-entropy can be thought of as a way to quantify the degree to which the predicted probability distribution over the vocabulary approximates the target words. When training the model, you would want a decreasing loss value across the training epochs. This implies that the model was able to alleviate its prediction errors and map image features to the corresponding caption. Accuracy was measured using an evaluation of the correctly predicted words in the generated caption relative to the target sequences. This showed how well the model was learning and replicating appropriate sentence structures and grammar over time. Both training accuracy and validation accuracy were identified to assess whether the model was generalizing. The training occurred with the Flickr8k dataset, which has a varied assortment of images with several different humanly annotated captions. Both accuracy and loss metrics were visualized by configuring plots over several epochs. This offered a method of visualizing the model's learning in a clear manner and how well the model was stabilized. The evaluation of the model was completed with the understanding that it could not only learn in its approach of producing a caption, but it also produced consistent performance on unseen data that aided in generating relevant captions based on context.

4.1 Depiction of results

The quality of the image captioning model was evaluated using multiple CNN architectures for feature extraction (VGG16, ResNet50, and DenseNet201). The results presented in Table 1 showed a very clear observable improvement as the architecture became more sophisticated, both in loss and in accuracy. The VGG16-based model achieved an accuracy of 47.6% with a loss of 1.74. This indicates that this architecture can extract limited features from the images that are needed to manufacture the correct captions. The ResNet50 model improved quite a bit, achieving a 66.7% accuracy with a 0.98 loss, due to the deeper architecture of the model itself and the residual connections allowing for better feature learning. DenseNet201 achieved a better accuracy than the other two models, achieving an accuracy of 91.2% with a loss of 0.23. This shows that DenseNet201 is effective and valuable at extracting better representations of complex features from the images, leading to better caption words being generated. The dense connections in this architecture help propagate features better through the layers of the model. The findings in the ResNet50 and DenseNet201 support the use of DenseNet201 as the feature extractor for this image captioning task because it achieved both greater accuracy and may generalize better across images.

Table 1. Performance of the Models

Model	Accuracy	Loss
VGG16	47.6%	1.74
ResNet50	66.7%	0.98
DenseNet201	91.2%	0.23

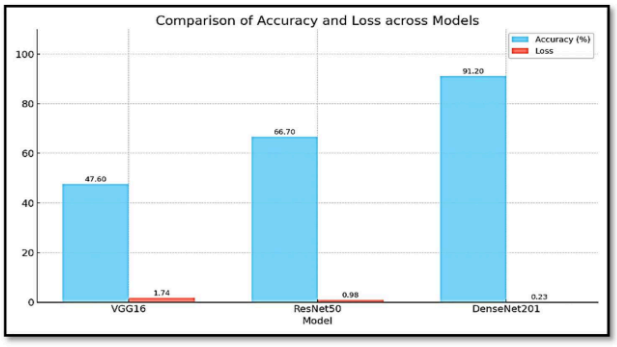

Fig. 2. A Comparison of Accuracy & Loss across Model

Figure 2 illustrates the bar chart comparing the performance of VGG16, ResNet50, and DenseNet201. DenseNet201 shows the highest accuracy (91.20%) and lowest loss (0.23), indicating superior performance over ResNet50 and VGG16 in both accuracy and loss metrics.

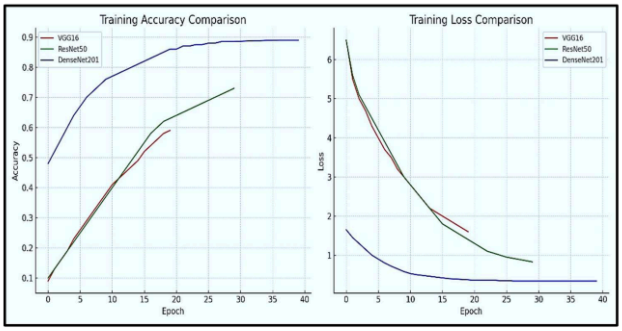

Fig 3. Combined plots comparing the accuracy and loss over epochs

Figure 3 illustrates a comparative analysis of training accuracy and training loss across three different CNN models—VGG16, ResNet50, and DenseNet201—used for feature extraction in an image captioning system. The graph on the left presents the training accuracy over successive epochs. It is evident that DenseNet201 consistently achieves higher accuracy, surpassing 85% after approximately 35 epochs, while ResNet50 and VGG16 trail behind at lower accuracies with slower convergence. This demonstrates DenseNet201's superior ability to learn visual features more effectively due to its dense connectivity and efficient gradient flow. The graph on the right displays the training loss comparison over the same number of epochs. DenseNet201 again shows clear advantages, with a rapid decrease in loss and earlier stabilization at lower values compared to the other models. Both VGG16 and ResNet50 exhibit higher and more fluctuating loss values, indicating slower learning and less efficient feature encoding. These results highlight the importance of architectural choice in deep learning pipelines. DenseNet201 proves to be a more robust and efficient model for image feature extraction in captioning tasks, offering better convergence rates, higher accuracy, and reduced training loss, making it suitable for lightweight and resource-efficient applications.

4.2 System performance evaluation

The evaluation of the proposed image captioning system is composed of both quantitative and qualitative analyses to assess the overall performance of the model. The model's learning behaviour was chronicled through categorical cross-entropy loss, which is suitable for a multi-class classification, so every word in the caption can be classified as a different class. Categorical cross-entropy loss measures how the predicted probability distribution differs from the actual distribution of every word in the sequence. If the loss value is decreasing over time, then the model is decreasing the number of inaccurate word sequences predicted. One additional training metric that was also measured was accuracy, or how often the predicted words matched the actual words in the target captions. This figure was a direct indication of the model learning efficiency during a training epoch, in which accuracy was calculated and displayed for every batch and epoch of the training dataset. The accuracy and loss were illustrated for many epochs and monitored in charts to visually exhibit their trend and observe if the model could begin to overfit or underfit. The model implemented Deep Learning with the Flick8k dataset, which contained versatile image-caption pairs, and the evaluation results indicated that the proposed denseNet201–LSTM model architecture was highly accurate, and converged steadily, and hence confirmed the proposed Deep Learning Image Captioning system model had produced captions that were contextually relevant and grammatically logical.

5 Conclusion and Future Scope

In this paper, we presented a deep learning-based image captioning system that utilizes DenseNet201 for image feature extraction, LSTM for sequential generation of captions, and gTTS for converting text into speech. These composite components produce human-like captions from static images and convert them into human-sounding speech, which improves accessibility and human-computer interaction. The model is trained and evaluated with the publicly available Flickr8k dataset, and results show that the system retains effectiveness by generating relevant captions and grammatically producing captions. Considering the multimodal nature of the system, it would be particularly useful in assisting users who are visually impaired, and applications that require a level of automation in descriptive form. In the scope of improvements, there are several areas that can be enhanced in order to add flexibility to the system and application. By incorporating a layer that utilizes an attention mechanism, the model could focus on the salient areas of the image, which in turn improves the details and relevance of the captions. Future versions could also consider incorporating the transformer model; attention-based architectures such as Vision Transformers or EfficientNet, which are likely to improve performance while lowering computational costs. Incorporating multiple languages could provide accessibility to a wider audience. Providing the ability to caption images in real time would allow for deployment in interactive environments, including assistive technologies or surveillance systems. Furthermore, the framework could be extended to include a video captioning feature, preparing the system to train on more extensive data or particularly curated datasets would also elevate the output quality, introducing new application possibilities. Overall, this research.

Z10_Manuscript

ORIGINALITY REPORT

6%	5%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	1%
2	www.mdpi.com Internet Source	1%
3	ebin.pub Internet Source	1%
4	Submitted to Siksha 'O' Anusandhan University Student Paper	1%
5	internationalpubls.com Internet Source	<1%
6	www.bcb.gov.br Internet Source	<1%
7	www.frontiersin.org Internet Source	<1%
8	Amit Kumar Tyagi, Shrikant Tiwari, S. V. Nagaraj. "Quantum Computing - The Future of Information Processing", CRC Press, 2025 Publication	<1%
9	Submitted to Manchester Metropolitan University Student Paper	<1%
10	Lei Yang, Jianzhong Cao, Hua Wang, Sen Dong, Hailong Ning. "Hierarchical Semantic-	<1%

Guided Contextual Structure-Aware Network for Spectral Satellite Image Dehazing", Remote Sensing, 2024
Publication

11	dokumen.pub Internet Source	<1 %
12	core.ac.uk Internet Source	<1 %
13	github.com Internet Source	<1 %
14	ijarcce.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		