NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

(ISO 9001:2015 Certified), Accredited with 'A' Grade by
NAAC 08258 - 281039 – 281263, Fax: 08258 – 281265

Department of Computer Science and Engineering

Artificial Intelligence

MINI PROJECT

On

Image To Voice Caption Generator

Course Code: 20CSE71

Academic Year – 2022-2023

Semester: 5 Section: C

Submitted To,

Course Instructor:

Ms. Shwetha G K
Assistant Professor-II
Department of CSE,
NMAMIT, Nitte.

Submitted By:

Name: Raksha Kamath USN: 4NM20CS143

Name: Shaina Crasta USN: 4NM20CS160

Date of submission: 13-12-2022

Signature of Faculty

Abstract

Making a computer system detect objects and describe them using natural language processing (NLP) in an age-old problem of Artificial Intelligence. This was considered an impossible task by computer vision researchers till now. With the growing advancements in Deep learning techniques, availability of vast datasets, and computational power, models are often built which will generate captions for an image. Image caption generation is a task that involves image processing and natural language processing concepts to recognize the context of an image and describe them in a natural language like English or any other language.

In this project, we use CNN and LSTM to identify the caption of the image. As the deep learning techniques are growing, huge datasets and computer power are helpful to build models that can generate captions for an image. This is what we are going to implement in this Python based project where we will use deep learning techniques like CNN and RNN. Image caption generator is a process which involves natural language processing and computer vision concepts to recognize the context of an image and present it in English.We used Keras library, numpy, Kaggle and flickr_dataset for the making of this project.

ACKNOWLEDGEMENT

We would like to give our sincere acknowledgement to everybody responsible for the successful completion of our project titled " Image To Voice Caption Generator ".

It gives us great pleasure to acknowledge with thanks the assistance and contribution of many individuals who have been actively involved at various stages of this project to make it a success. Firstly, we are very grateful to this esteemed institute "NMAM Institute of Technology " for providing us an opportunity for our degree course. We wish to express our whole hearted thanks to our principal DR. NIRANJAN.N.CHIPLUNKAR for providing the modernized lab facilities in our institute. Success does not happen overnight, but hard work and dedication can achieve any goal. We have tried our level best to fulfill the requirements of the project, but I could not have achieved my goal without the able guidance of Ms SHWETHA, Assistant Professor. We thank our guide for providing us an opportunity to work under their guidance and for their constant support and encouragement.

We thank one and all for helping us during our project .

Table of Contents

Introduction

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions. However, machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. Captions for every image on the internet can lead to faster and descriptively accurate images searches and indexing. Ever since researchers started working on object recognition in images, it became clear that only providing the names of the objects recognized does not make such a good impression as a full human-like description. As long as machines do not think, talk, and behave like humans, natural language descriptions will remain a challenge to be solved.

Literature Review

Sl.no	Topic	Journal Publication Year	Objectives	Methods	Scope of Improvement/ Challenges	Conclusion
1	Image caption generator using image features and lstm Networks	July-2021	The development of image description started with object detection using static object class libraries in the image and modeled using statistical language models.	The image preprocessing is carried out by feeding the input data to the Xception application of the Keras API, which is running on top of TensorFlow, and the dataset utilized is Flickr 8K. CNN + LSTM is to take an image as input and produce a caption.	As only at the beginning of the process the picture information given, hence vanishing gradient issues could arise. As a result, creating extended sentences with LSTM is difficult.	Together, the CNN and LSTM were able to determine the relationship between objects in images by synchronizing their operations and were able to produce hybrid image caption generators as a result.

2	A Hybrid Model for Combining Neural Image Caption and k-Nearest Neighbor Approach for Image Captioning	Aug-2021	This suggested hybrid model incorporates two cutting-edge models: Neural Image Caption and k-Nearest Neighbor method. The logistic regression is applied in this hybrid technique to select the best model for a particular input image and produce captions.	Let M1 and M2 be the NIC model and the k-nearest neighbor model , respectively that are trained individually on the training set of images and the captions. A hybrid model M which selects a consensus caption c* for a given input image I.	Bias in the dataset has an effect that causes trained models to overfit on similar or common items, making it difficult to generalize the appropriate description. In order to bridge the meaning gap between language and visual, common sense and logic must be incorporated.	The method suggests combining two image captioning models and training a logistic regression classifier. The system has attained superior BLEU-1 and BLUE-4 scores on Flickr8k dataset. This method can be expanded to incorporate more than two image captioning models

3	Image Caption Generating Deep Learning Model	Sept-2021	Neural networks concept is used in this paper to caption the images. With the aid of built-in vocabulary and image features, CNN (ResNet) is utilised as an encoder to access the image features, and RNN (Long Short Term Memory) is used as a decoder to provide captions for the images.	CNN-RNN based framework, where RNN is used for the decoding process and CNN is used for encoding. utilising the NLTK library, and mapped with Image characteristics to obtain the precise word for the supplied image.	While CNN-RNN based captions may have less loss than CNN-CNN based captions, the training period is longer. Training time has an impact on the model's overall efficiency.	ResNet-LSTM model has higher accuracy compared to simple CNN-RNN and VGG Model and beneficial for processing massive volumes of unstructured and unlabeled data to detect patterns in photos

4	Image captioning generator system with Caption to speech conversion mechanism	Jan-2021	Provide the ideal caption for a picture. Which requires perfect recognition of the scenes and objects provided, and also needs the ability to analyze the state with the relationships and attributes in the objects obtained. And then it is converted into voice speech.	This model was developed using flickr8k and ImageNet datasets. Using the InceptionV3 model, transfer learning was used to convert the images into fixed size vectors (CNN). RNN is used to process the sequence and the data matrix also in partial captions using the LSTM layer. gTTS is used to voice-speak the generated caption.	Auto-regressive based models suffer from slow inference problem in gTTS	The system uses the image vectors to create partial caption vectors, merges them, and then creates the best captions for the images. The captions are also converted to speech to assist those who are blind or visually impaired.

5	Automatic Image Caption Generation Using Deep Learning	June-2022	A deep neural network-based approach to creating image captions is suggested. This method uses a convolutional neural network (VGG16 Hybrid Places1365) as the encoder and a recurrent neural network (LSTM) as the decoder to create captions using a deep neural network model.	CNN is used as encoders(VGG 16 Hybrid Places 1365) and for sentence generation RNN (LSTM) is used. In order to measure the performance of the model,the evaluation process verifies the different characteristics of the generated captions such as readability, grammatical of the caption using evaluation metrics like BLEU, GLEU on Flickr8k and MSCOCO Captions datasets.	Crucial downsides of the VGG16 network is that it is a huge network, which means that it takes more time to train its parameters. This makes implementing a VGG network a time-consuming task.	Study proposed an encoder-decoder based model to generate grammatically correct captions for images. The experimental results show that the model surpasses all existing state-of-the art scores. The paper also reported results of caption generation on live sample images that reinforces the validity of the proposed approach.

Requirements

- Modules and packages:
    - ResNet-50 model
    - Tensorflow
    - Keras

- Programming Language: Python

- Editor used: Kaggle

Design

The planned project design will involve the below listed functionalities:

- A frontend interface which would be allow users to
  - Upload an image for caption generation.
  - Download the caption generated in audio format.
- A backend which would be involved in
  - Extracting the features from the images for caption generation and storing them.
  - Generation of captions for the uploaded images.
  - Image and Caption dataset which is used to train the model.
  - Conversion of generated caption into audio format.

Implementation

Frontend design breakdown:

- Users upload images for which they want a caption to be generated.
- Users get a detailed description of the image which can be copied.
- The trained model generates the caption for the uploaded image.
- Users also have the option to download the audio version on the respectively generated caption.

Backend design breakdown:

- For this project we are using keras on Kaggle and Resnet50 pretrained model, which is already available in keras, makes it possible to train the network on thousands of layers without degrading performance.
- The dataset which we are using is Flickr8k.

Flickr8k_Dataset: Contains a total of 8092 images in JPEG format with different shapes and sizes. Of which 6000 are used for training, 1000 for test and 1000 for development.

Flickr8k_text : We are using Flickr8k.token.txt which has 5 captions for each image.

- We use glob to read the .jpg images from the given directory and with the help of cv2 the images are resized and reshaped.

1. Preprocessing of the images : It is done by using Flickr dataset with the help of Resnet50. We extract the features of the image which will be converted into vectors of 2048 values by removing the dense layer (last layer in resnet) and store it in a dictionary. Due to due to insufficiency of ram, only 1500 images can be processed from flicker8k dataset.

2. Preprocessing of captions: Every caption is given sos(start of string) and eos (end of string) token to the captions. Then they are split at spaces and with the help of to_categorical and pad_sequences the first word will be taken as input for our model and it predicts the next word.

3. Visualize Images with captions: In this step we define a function which will load all the words in that caption and add keys that is start of string token (sos) and end of string token (eos) , with the help of key and value we iterate through the dictionary to which it was previously added to.

4. Create Vocabulary

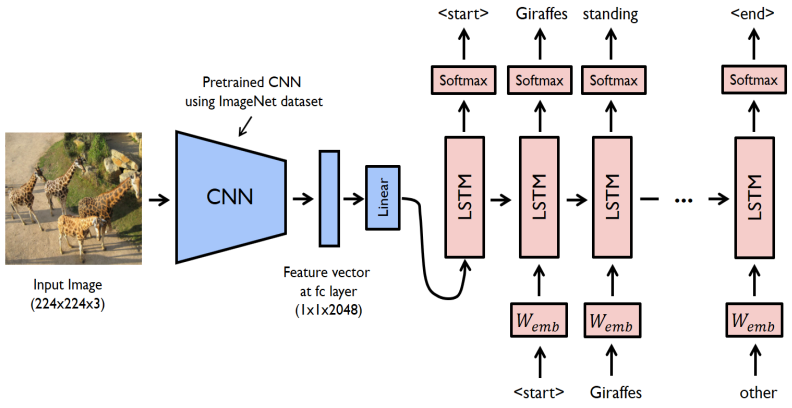
After finding the number of words present in dictionary in giving integer values to it and we convert the caption string into integer and kept in a list .

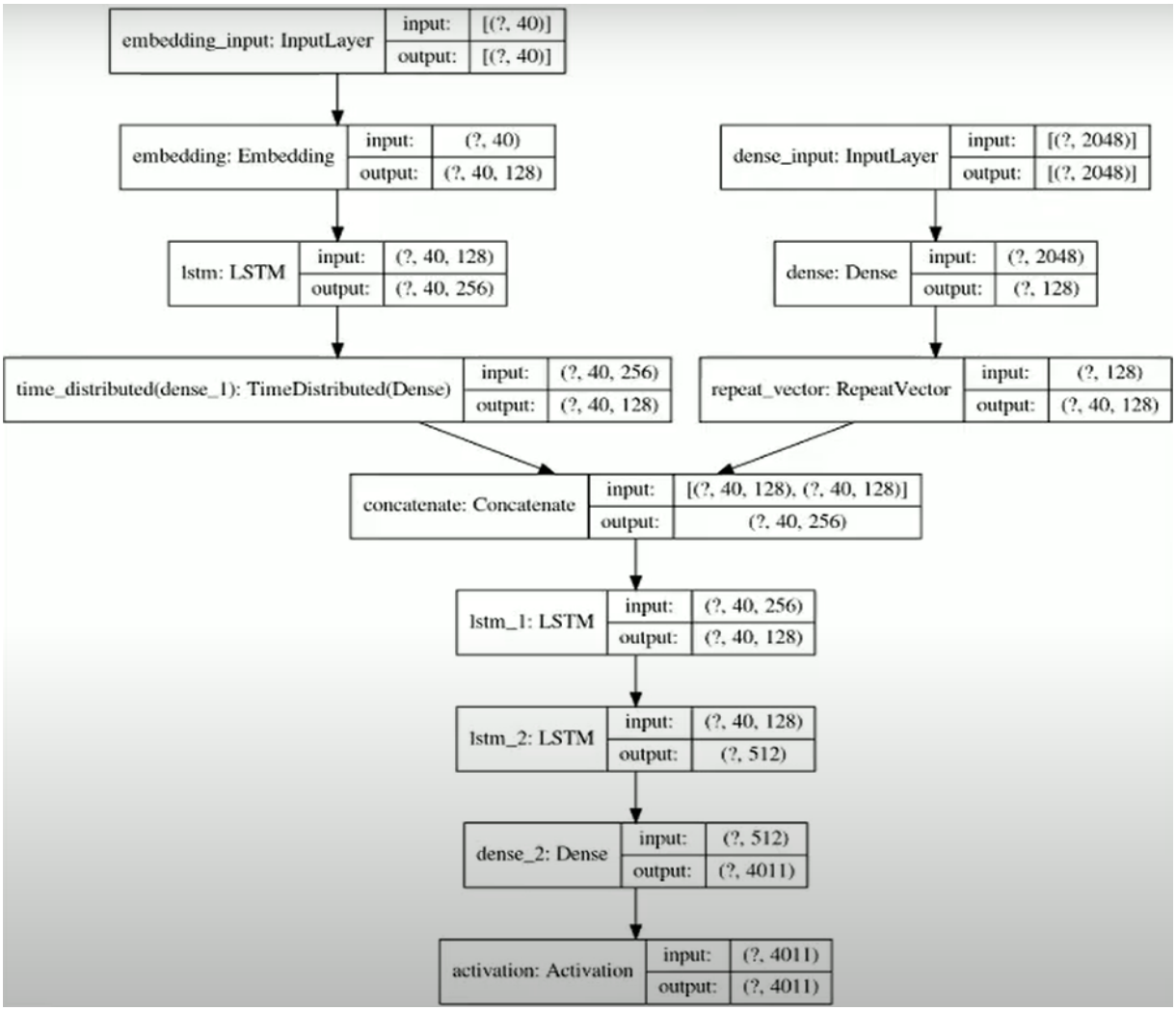
5. Build Generator Function

In this step we generate X, Y_in, Y_out by taking photos and catptions. Appending image features to the captions. The word with maximum probability will be appended to create a sentence where 1 is maximum probability and 0 is the least probability and stored as arrays so that the processing can be done faster.

6.Model

With the help of we initially process the caption with the size of 40 and the sample size which will finally give 40/128 beth size matrix. Next images is processed 2048 size is transferred into dense then its converted into 128 values and with the help of repete vector which has 40, 128 values in it. Now that images and captions are ready they are concatinated and transferred into LSTM in which in the Dense layer it has to check the probability of a word 4011 times and With the help of model.fit it produces y_out .





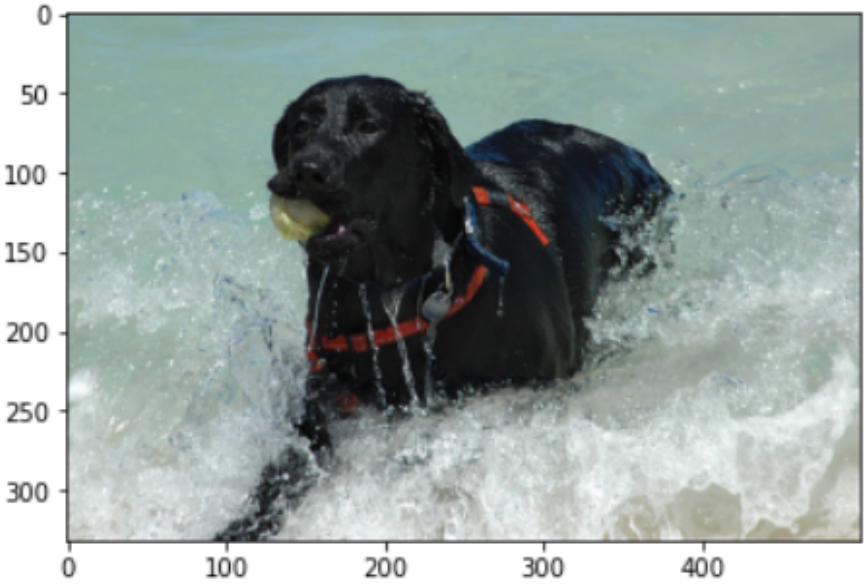
7. Predictions and Speech conversion : Now when the user uploads the image, it will be stored in a static folder, the saved models predict the caption, The predicted caption is then converted into speech with the help of Google text to speech library (gTTS) and downloadable mp3 format.

Results

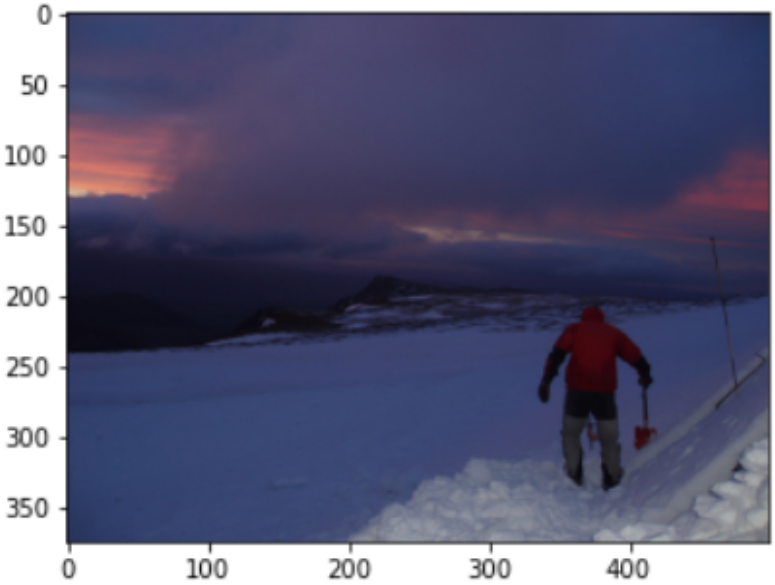
Training:

```
Epoch 43/50
188/188 [==============================] - 14s 72ms/step - loss: 1.2731 - accuracy: 0.6
799
Epoch 44/50
188/188 [==============================] - 14s 73ms/step - loss: 1.2375 - accuracy: 0.6
868
Epoch 45/50
188/188 [==============================] - 14s 75ms/step - loss: 1.2035 - accuracy: 0.6
946
Epoch 46/50
188/188 [==============================] - 13s 71ms/step - loss: 1.1727 - accuracy: 0.7
018
Epoch 47/50
188/188 [==============================] - 14s 73ms/step - loss: 1.1377 - accuracy: 0.7
108
Epoch 48/50
188/188 [==============================] - 14s 73ms/step - loss: 1.1055 - accuracy: 0.7
193
Epoch 49/50
188/188 [==============================] - 14s 73ms/step - loss: 1.0717 - accuracy: 0.7
266
Epoch 50/50
188/188 [==============================] - 14s 75ms/step - loss: 1.0449 - accuracy: 0.7
328
```

Testing phase:

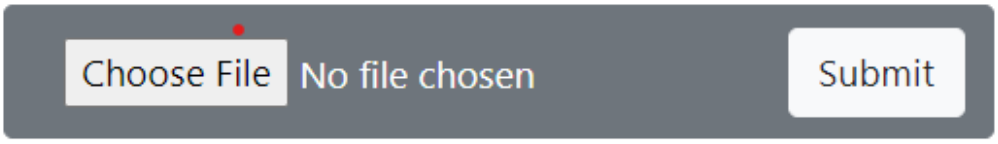


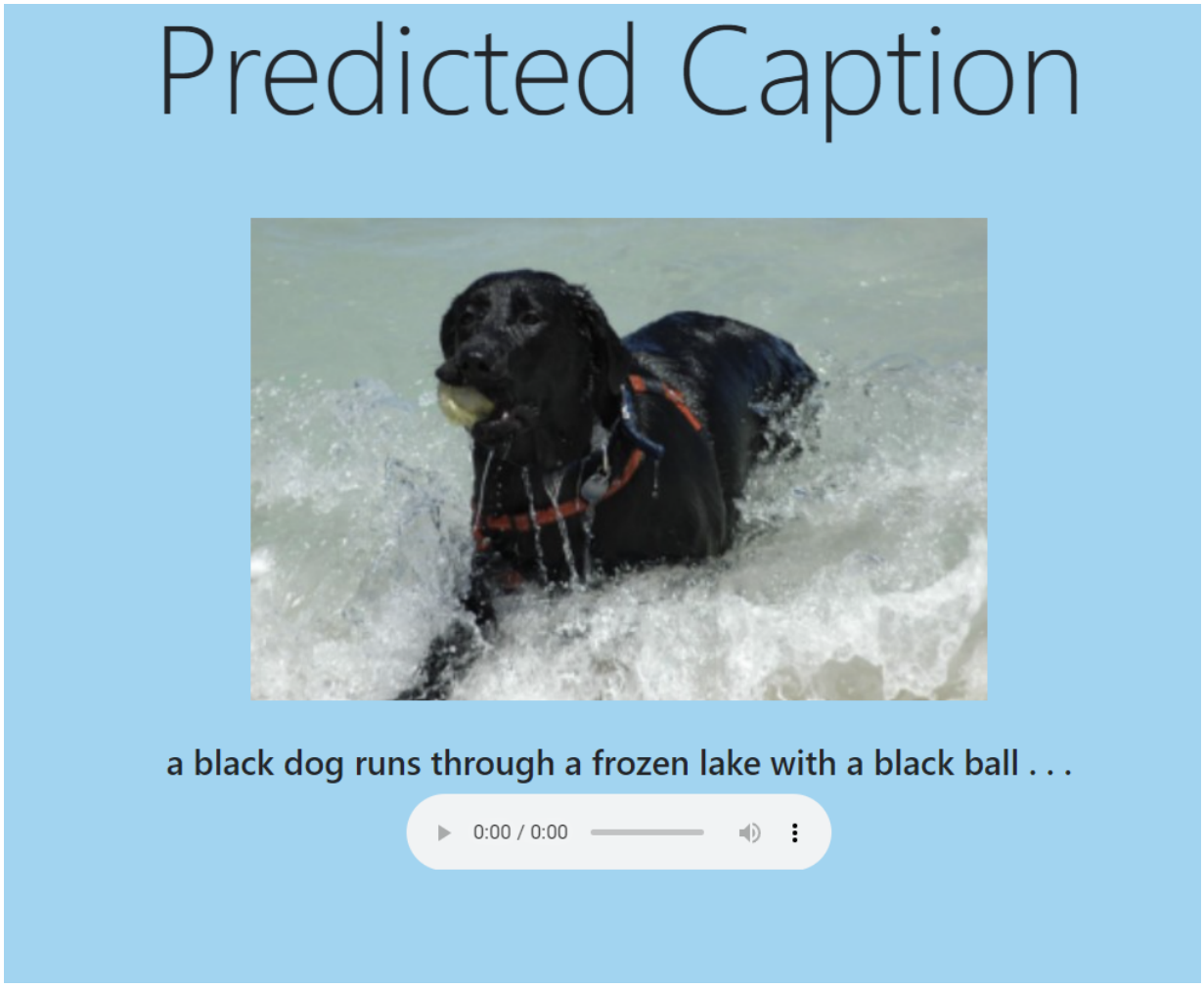
a black dog climbing out of the pool . endofseq



a hiker in a red jacket is carrying rocks and the arms . endofseq

Final Result:





Conclusion

Using deep learning-based image captioning methods and Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Although deep learning -based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for sometime. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

References

- https://github.com/aswintechguy/Deep-Learning-Projects/tree/main/Image%20Caption%20Generator%20-%20Flickr%20Dataset
- https://ijcseonline.org/pub_paper/3-IJCSE-08233.pdf
- https://ieeexplore.ieee.org/document/8728516/keywords#keywords