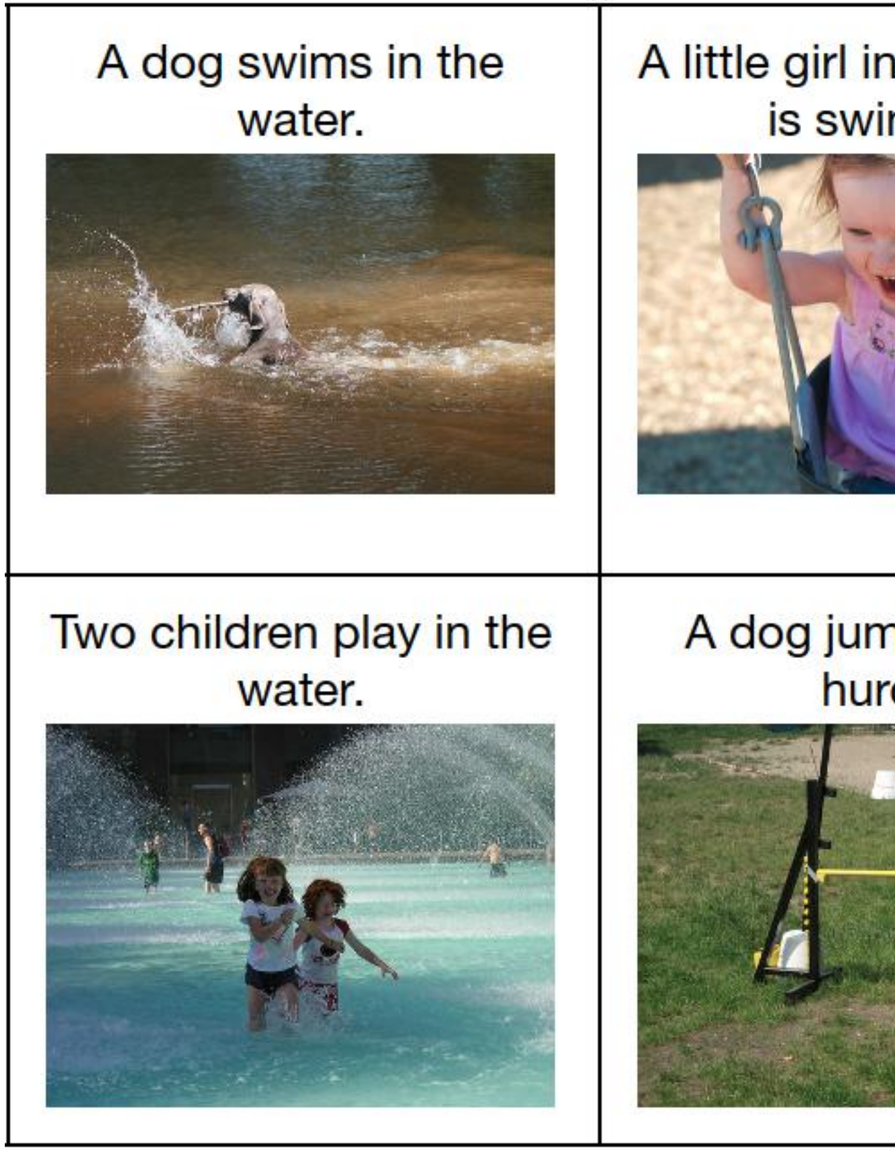
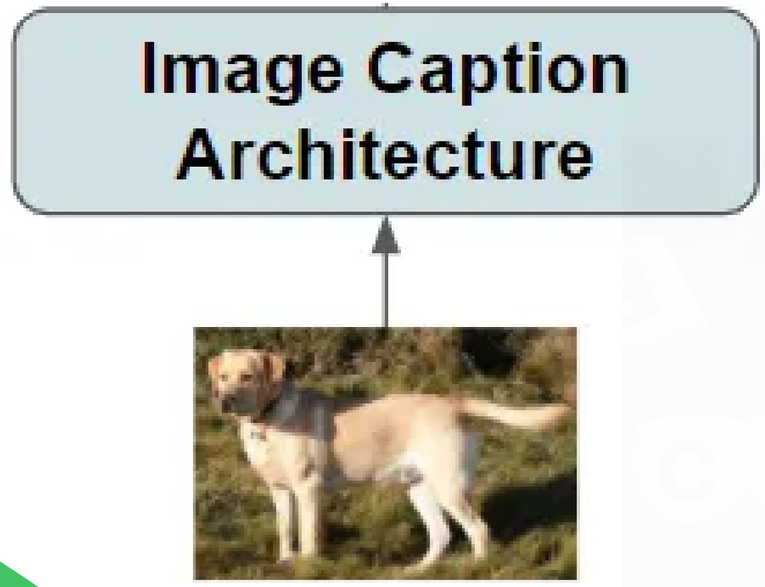
Captionator

Image Captioning Using Attention





Introduction

Generating a description of an image is called image captioning.

Image captioning requires recognizing relevant objects, their attributes, and relationships in an image. It also needs to generate syntactically and semantically correct sentences.

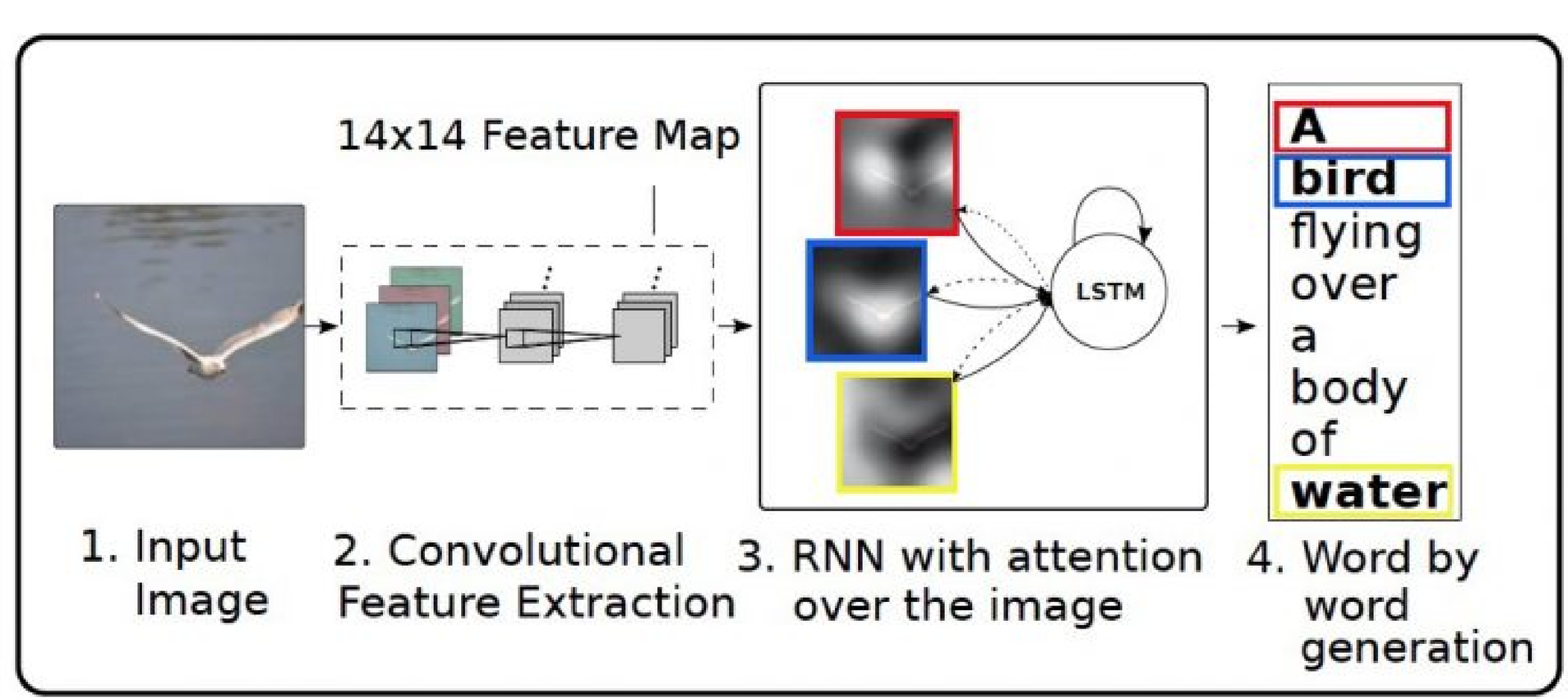
"Show, Attend and Tell" review

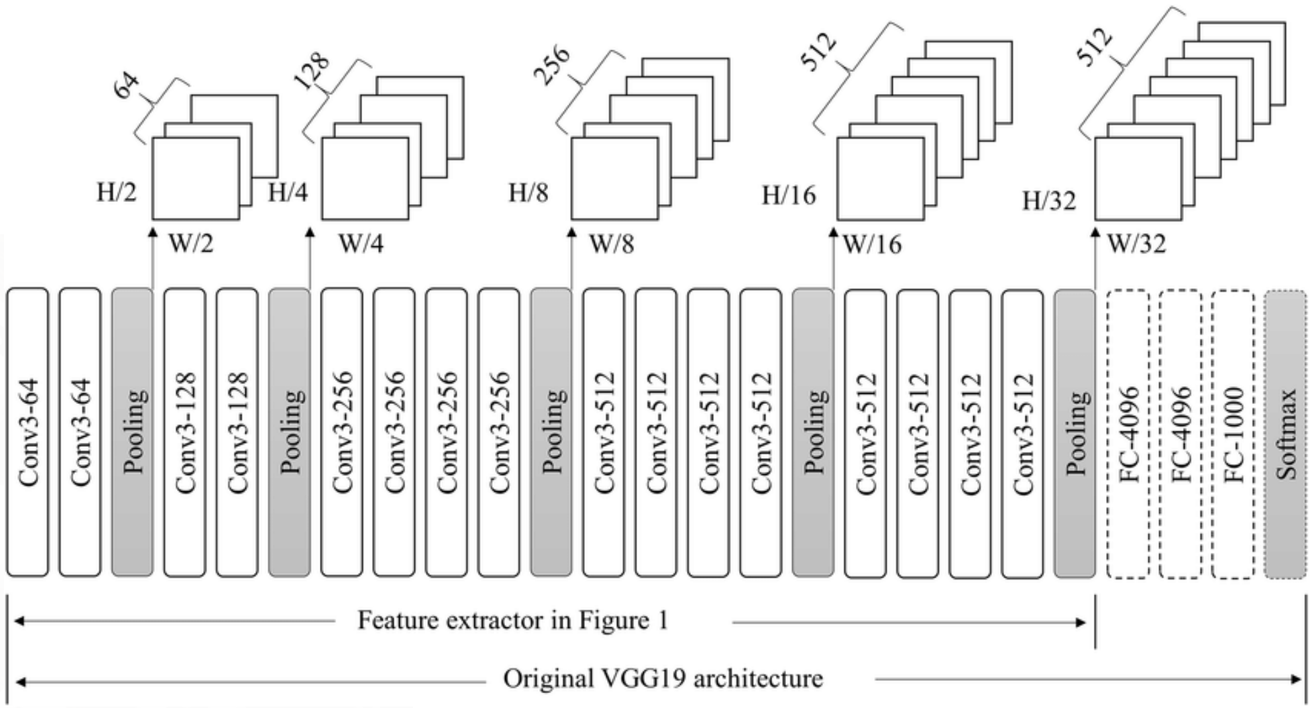
"Show, Attend and Tell" is a neural network model for image captioning, which was introduced in a research paper by Xu et al. in 2015. It uses a combination of convolutional neural networks (CNNs) to extract visual features from images and recurrent neural networks (RNNs) to generate captions for those images. The model is able to attend to different regions of the image as it generates captions, which improves the quality of the generated captions. The paper reported state-of-the-art performance on several benchmark datasets for image captioning.

"Neural Machine Translation by Jointly Learning to Align and Translate" review

It a research paper by Bahdanau et al. that proposes a neural machine translation model that learns to align and translate simultaneously. The model uses an attention mechanism that enables it to focus on different parts of the source sentence during the translation process, producing better translations compared to traditional machine translation models. The paper's approach has become a cornerstone of modern neural machine translation systems.

Approach to the problem statement





VGG-19

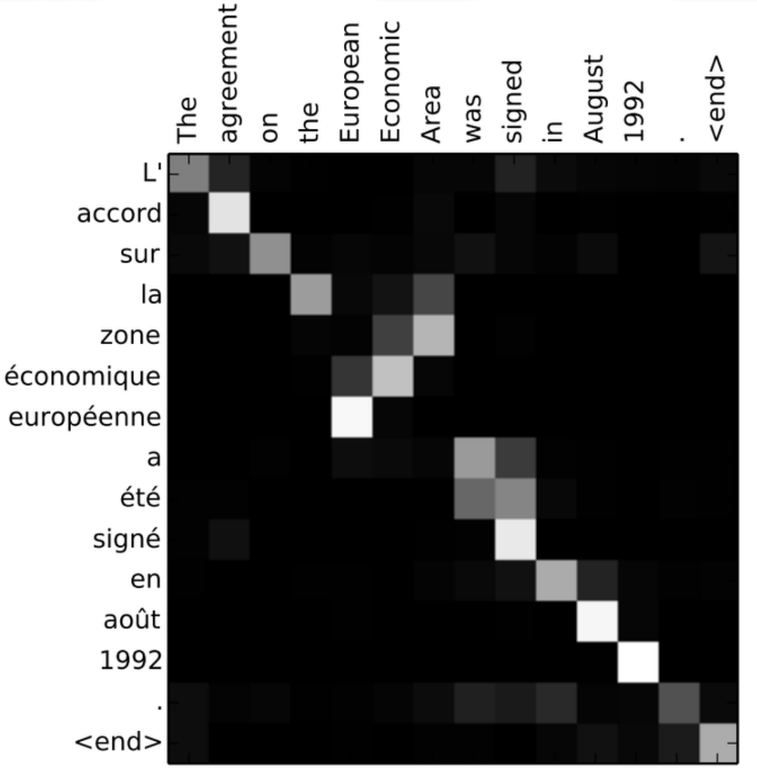
VGG19 is a deep convolutional neural network architecture that was developed for image recognition tasks.

It consists of 19 layers including convolutional, pooling, and fully connected layers. VGG19 is known for its high accuracy and is often used as a benchmark for image classification tasks.

Bahdanau Attention

Additive Attention, also known as Bahdanau Attention, uses a one-hidden layer feed-forward network to calculate the attention alignment score:

- Additive attention computes a set of attention scores by taking the weighted sum of a learned representation of the decoder's current hidden state and the encoder's output at each time step.
- Additive attention is a type of content-based attention as it focuses on the specific content of the input sequence to determine the importance of each input element at each decoding step.
- The attention scores are then used to compute a weighted sum of the encoder output, which is used as context information for the decoder at each decoding step.



Dataset

Flickr8k dataset in which each image is associated with five different captions that describe the entities and events depicted in the image that were collected.

Dataset structure is as follows:-

Flick8k/
    Flick8k_Dataset/ :- contains the 8091 images
Flick8k_Text/
    Flickr8k.token.txt:- contains the image id along with the 5 captions

Defined an 80:20 split to create training and test data.

Steps of learning