

PROJECT REPORT

(Project Term August-November 2021)

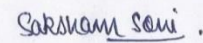
IMAGE CAPTION GENERATOR

Submitted by

Name of Student: Saksham Soni

Registration Number: 11913397

Signature of the Student:



Course Code: INT 246

Under the Guidance of

(Dr. Sagar Pande)

School of Computer Science and Engineering



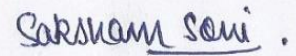
**L OVELY
P ROFESSIONAL
U NIVERSITY**

Declaration

I hereby declare that the project work entitled (“Image Caption Generator”) is an authentic record of our own work carried out as requirements of Project for the award of B.Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of (Dr. Sagar Pande), during August to November 2021. All the information furnished in this project report is based on our own intensive work and is genuine.

Name of the Student: Saksham Soni

Registration Number: 11913397

A rectangular box containing a handwritten signature in blue ink that reads "Saksham Soni".

Date: 05-11-21

CERTIFICATE

This is to certify that the declaration statement made by this student is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B.Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara.

Name of the Mentor: Dr. Sagar Pande

Designation

School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab.

ACKNOWLEDGEMENT

I would like to express my gratitude towards my University , my Mentor(Dr. Sagar Pande) and Udeemy for providing me guidance for this Project and Machine Learning, which also helped me in doing a lot of homework and learning. As a result, I came to know about so many new things. So, I am really thank full to them.

Moreover I would like to thank my friends who helped me a lot whenever I got stuck in some problem related to my course. I am really thankful to have such a good support of them as they always have my back whenever I need.

Also,I would like to mention the support system and consideration of my parents who have always been there in my life to make me choose right thing and oppose the wrong. Without them I could never had learned and became a person who I am now.

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them

TABLE OF CONTENTS

Title Page.....	(i)
Declaration.....	(ii)
Certificate.....	(iii)
Acknowledgement.....	(iv)
Table of Contents.....	(v)

1. INTRODUCTION OF THE PROJECT UNDERTAKEN

2. SCOPE OF THE PROJECT AND USE CASE

3. VARIOUS CONCEPTS AND FRAMEWORKS USED

- **Numpy**
- **Pandas**
- **Supervised Learning**
- **Neural Networks**
- **Deep Learning**
- **CNN**
- **RNN**
- **Transfer Learning**
- **Word Embeddings**

4. ABOUT DATASET AND SOURCE CODE

5. CONCLUSION

INTRODUCTION OF THE PROJECT UNDERTAKEN

Objectives of the project

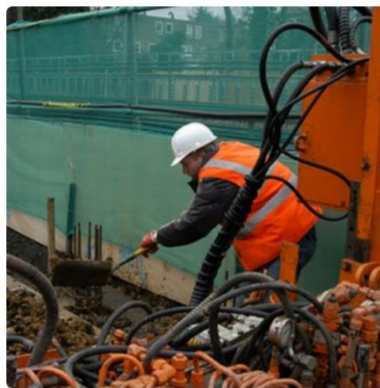
Machine Learning is a field of technology developing with immense abilities and applications in automating tasks, where neither human intervention is needed nor explicit programming.

The power of ML is such great that we can see its applications trending almost everywhere in our day-to-day lives. ML has solved many problems that existed earlier and have made businesses in the world progress to a great extent. Automatically generating captions to an image shows the understanding of the image by computers, which is a fundamental task of intelligence. For a caption model it not only need to find which objects are contained in the image and also need to be able to be expressing their relationships in a natural language such as English. Recently work also achieve the presence of attention, which can store and report the information and relationship between some most salient features and clusters in the image.

Image captioning can be regarded as an end-to-end Sequence to Sequence problem, as it converts images, which is regarded as a sequence of pixels to a sequence of words. For this purpose, we need to process both the language or statements and the images. For the Language part, we use recurrent Neural Networks and for the Image part, we use Convolutional Neural Networks to obtain the feature vectors respectively.



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



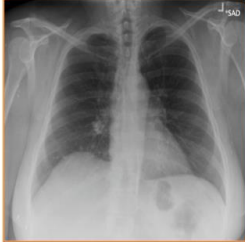





Scope of the project

Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain. Image captioning has a huge amount of application. NVIDIA is using image captioning technologies to create an application to help people who have low or no eyesight.

In our project, we do image-to-sentence generation. This application bridges vision and natural language. If we can do well in this task, we can then utilize natural language processing technologies understand the world in images. In addition, we introduced attention mechanism, which is able to recognize what a word refers to in the image, and thus summarize the relationship between objects in the image. This will be a powerful tool to utilize the massive unformatted image data, which dominate the whole data in the world.

Use Cases

- Some detailed use cases would be like a visually impaired person taking a picture from his phone and then the caption generator will turn the caption to speech for him to understand.
- Advertising industry trying the generate captions automatically without the need to make them separately during production and sales.
- Doctors can use this technology to find tumors or some defects in the images or used by people for understanding geospatial images where they can find out more details about the terrain

input image				
	aorta_thoracic / tortuous / mild aorta_thoracic / tortuous	opacity / lung / middle_lobe / right / aorta_thoracic / tortuous opacity / lung / base / left	calcified_granuloma / lung / middle_lobe / right / multiple calcified_granuloma / lung / hilum / right	opacity / lung / middle_lobe / right / blood_vessels calcified_granuloma / lung / middle_lobe / right
generated annotation				
true annotation				
	airspace_disease / lung / hilum / right / lung / hilum nodule / lung / hilum / right	thoracic_vertebrae_degenerative / mild aorta_tortuous / thoracic_vertebrae_degenerative / mild	normal normal	normal normal

HOW IMAGE CAPTIONING WORKS



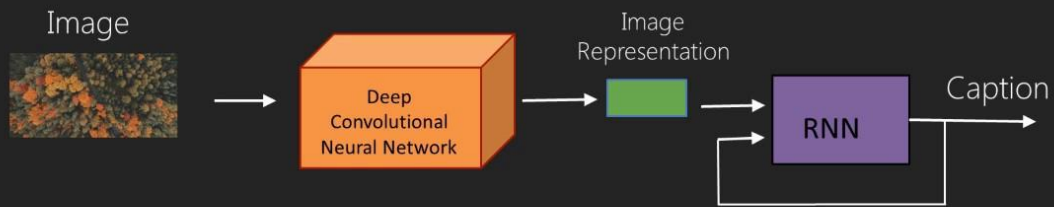
2.1

If we are told to describe it, maybe we will describe it as: “A puppy on a blue towel” or “A brown dog playing with a green ball”. So, how are we doing this? While forming the description, we are seeing the image but at the same time, we are looking to create a meaningful sequence of words. The first part is handled by CNNs and the second is handled by RNNs.

The task of image captioning can be divided into two modules logically – one is an **image based model** – which extracts the features and nuances out of our image, and the other is a **language based model** – which translates the features and objects given by our image based model to a natural sentence.

For our image-based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language-based model (viz decoder) – we rely on a Recurrent Neural Network. The image below summarizes the approach given above

Automatic Image Caption



2.2

Usually, a pretrained CNN extracts the features from our input image. The feature vector is linearly transformed to have the same dimension as the input dimension of the RNN/LSTM network. This network is trained as a language model on our feature vector. Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short-Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords.

VARIOUS CONCEPTS AND FRAMEWORKS USED

NUMPY

The NumPy array is a data structure that efficiently stores and accesses multidimensional arrays (also known as tensors) and enables a wide variety of scientific computation. It consists of a pointer to memory, along with metadata used to interpret the data stored there, notably 'data type', 'shape' and 'strides'



3.1

PANDAS

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allows us to analyze big data and make conclusions based on statistical theories.

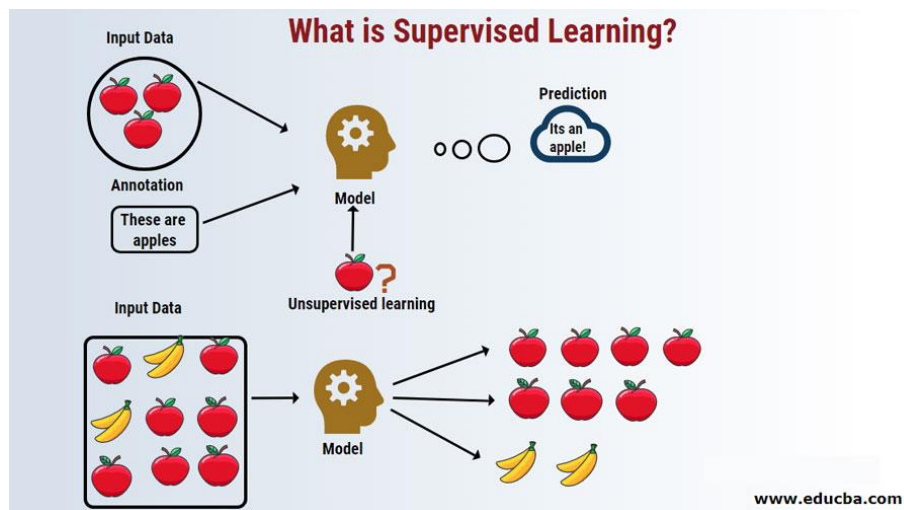
Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.



3.2

SUPERVISED LEARNING

A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalizes to respond correctly to all possible inputs. This is also called learning from exemplars. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. In supervised learning, each example in the training set is a pair consisting of an input object (typically a vector) and an output value. A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples. In the optimal case, the function will correctly determine the class labels for unseen instances. Both classification and regression problems are supervised learning problems. A wide range of supervised learning algorithms are available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems



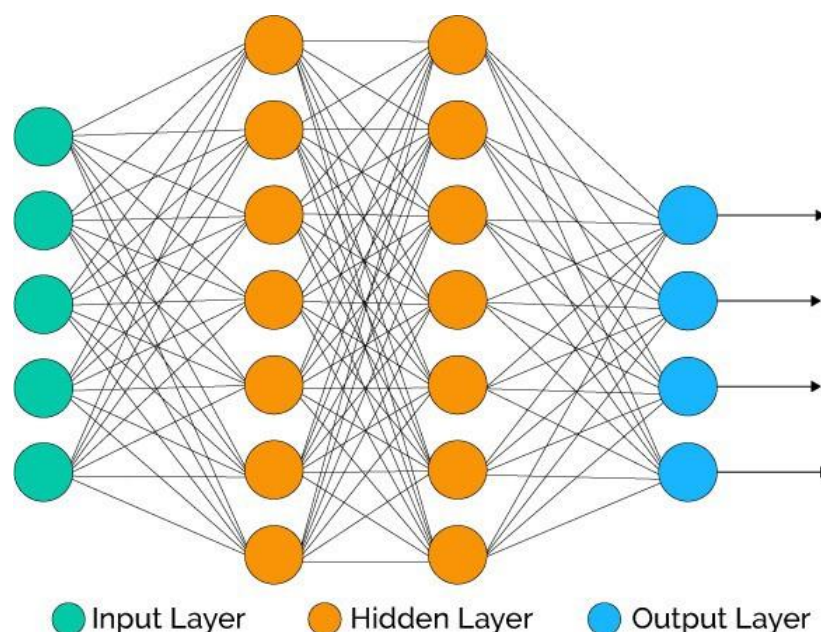
3.3

NEURAL NETWORKS

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layer, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm.

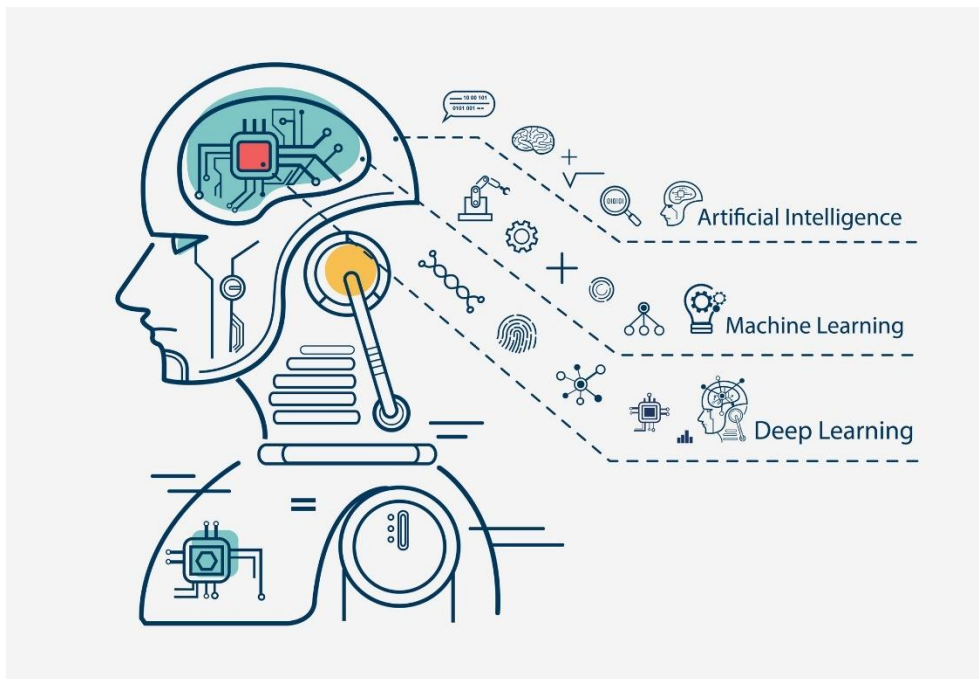


DEEP LEARNING

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.

Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services (such as digital assistants, voice-enabled TV remotes, and credit card fraud detection) as well as emerging technologies (such as self-driving cars).

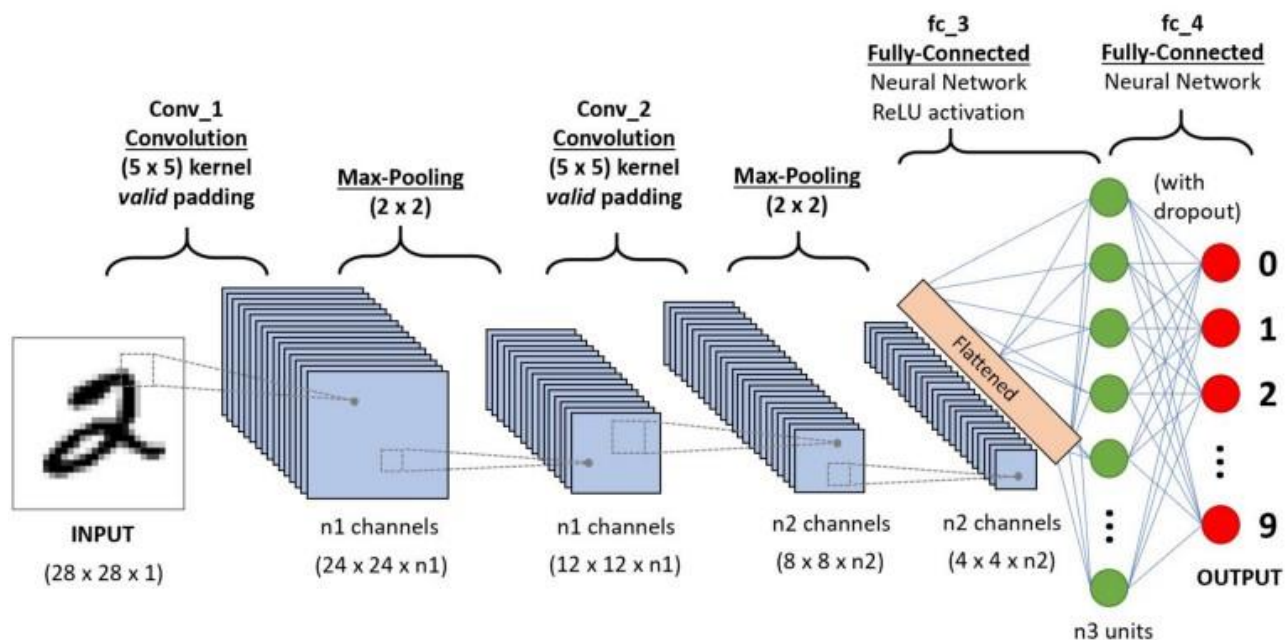
Deep learning neural networks, or artificial neural networks, attempts to mimic the human brain through a combination of data inputs, weights, and bias. These elements work together to accurately recognize, classify, and describe objects within the data.



CONVOLUTIONAL NEURAL NETWORK (CNN)

A **Convolutional Neural Network (CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics.

The architecture of a CNN is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlaps to cover the entire visual area.

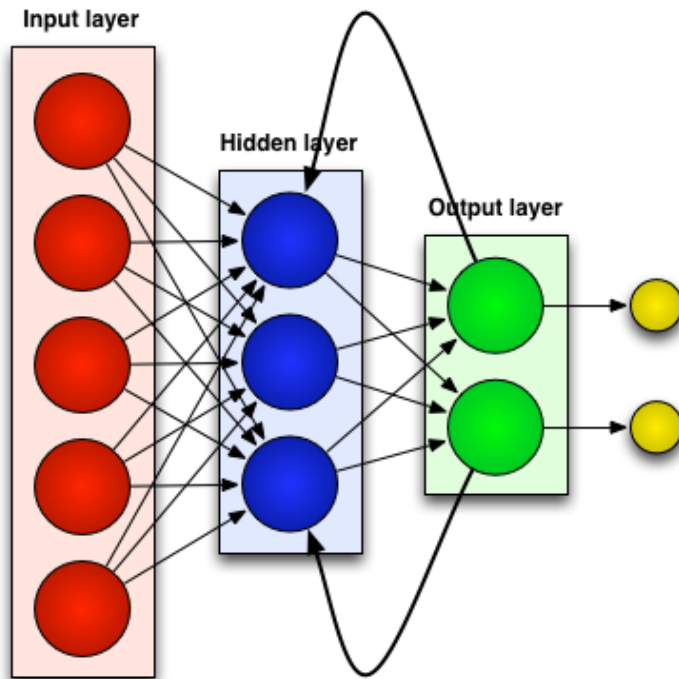


3.6

RECURRENT NEURAL NETWORK (RNN)

A recurrent neural network (RNN) is a special type of an artificial neural network adapted to work for time series data or data that involves sequences. Ordinary feed forward neural networks are only meant for data points, which are independent of each other. However, if we have data in a sequence such that one data point depends upon the previous data point, we need to modify the neural network to incorporate the dependencies between these data points. RNNs have the concept of 'memory' that helps them store the states or information of previous inputs to generate the next output of the sequence.

RNN is the extension of feedforward NN with the presence of loops in hidden layers. RNN takes the input with the sequence of samples and identifies the time relationship between the samples. The Long short-term memory (LSTM) solves the classification issues by adding the network parameters with the hidden node and releases the state based on the input values. RNN achieves better performance than LSTM by activating the states based on network events. The regular RNN node consists of a single bias and weight. The RNN is evaluated using the gated recurrent unit and LSTM.



3.7

TRANSFER LEARNING

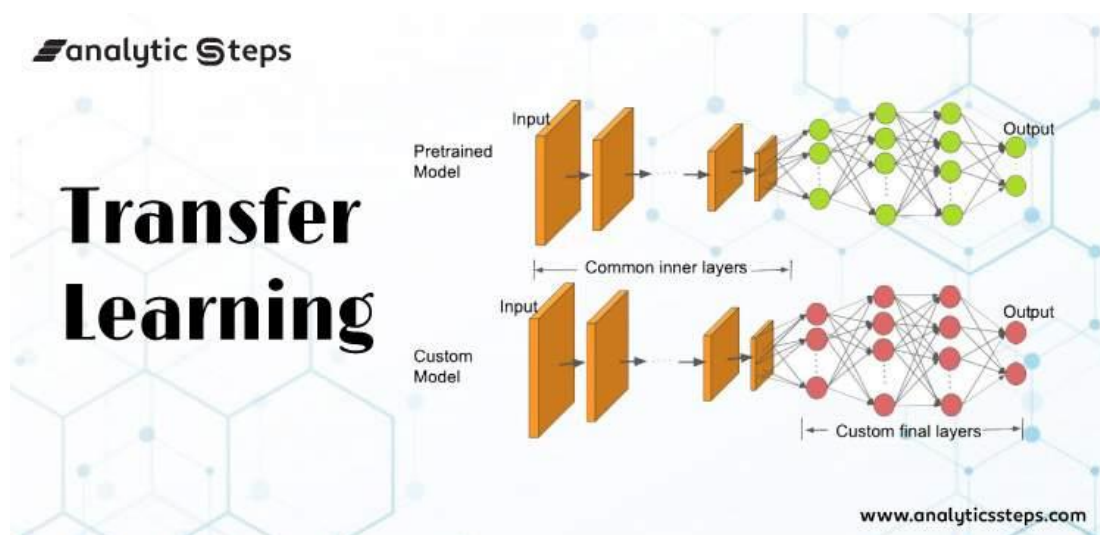
Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.

It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

In this post, you will discover how you can use transfer learning to speed up training and improve the performance of your deep learning model

Features of Transfer Learning

- It involves the transfer of knowledge that is grasped in one source task to learn and refine the related target task.
- It has been observed that DNN trained on the natural images shows a strange occurrence where the first layer of the network appears to learn features alike to Gabor filters.
- These first layers features are found to be general features for many datasets.
- Those features in the first layers ignoring the image dataset, task, loss function, are considered general features.

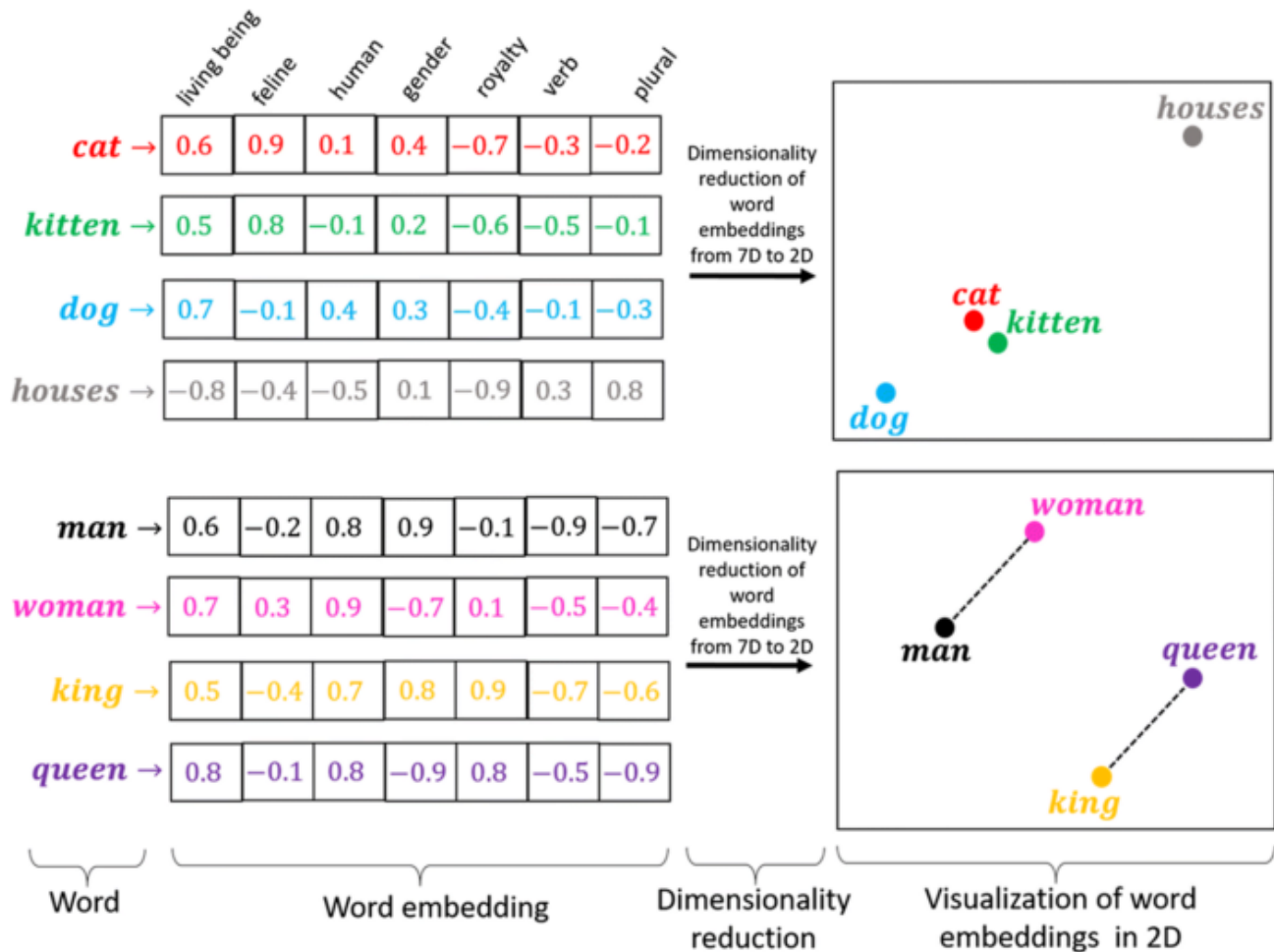


WORD EMBEDDINGS

A word embedding is a learned representation for text where words that have the same meaning have a similar representation.

It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems.

Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are learned in a way that resembles a neural network, and hence the technique is often lumped into the field of deep learning



ABOUT DATA SET AND SOURCE CODE

I have used flickr8k data set in this project which is available on kaggle website. There are various others versions for the same dataset are also available.

Link of the dataset: <https://www.kaggle.com/shadabhussain/flickr8k>

Steps I followed to build this project:

- Data collection
- Understanding the data
- Data Cleaning
- Loading the training set
- Data Preprocessing — Images
- Data Preprocessing — Captions
- Data Preparation using Generator Function
- Word Embeddings
- Model Architecture
- Inference

I have uploaded my project on GitHub.

Link of project on GitHub: <https://github.com/SakshamSoni-code/Image-Caption-Generator>

CONCLUSION

Automatically image captioning is far from mature and there are a lot of ongoing research projects aiming for more accurate image feature extraction and semantically better sentence generation. We successfully completed what we mentioned in the project proposal but used a smaller dataset (Flickr8k) due to limited computational power. There can be potential improvements if given more time. First of all, we directly used pre-trained CNN network as part of our pipeline without fine-tuning, so the network does not adapt to this specific training dataset. Thus, by experimenting with different CNN pre-trained networks and enabling fine-tuning, we expect to achieve a slightly higher BLEU4 score. Another potential improvement is by training on a combination of Flickr8k, Flickr30k, and MSCOCO. In general, the more diverse training dataset the network has seen, the more accurate the output will be. We all agree this project ignites our interest in application of Machine Learning knowledge in Computer Vision and expects to explore more in the future.