iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47058

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Image Caption Generator Using Deep Learning

Palak Kabra[1], Mihir Gharat[2], Dhiraj Jha[3], Shailesh Sangle[4]
[1, 2, 3, 4]*Department of Computer Engineering, Thakur College of Engineering and Technology*

Abstract: ***Nowadays, an image caption generator has become the need of the hour, be it for social media enthusiasts or visually impaired people. It can be used as a plugin in currently trending social media platforms to recommend suitable captions for people to attach to their post or can be used by visually impaired people to understand the image content on the web thus eradicating any ambiguity in image meaning in turn also free of any discrepancy in knowledge acquisition. The proposed paper aims to generate a description of an image also called as image captioning, using CNN-LSTM architecture such that CNN layers will help in extraction of the input data and LSTM will extract relevant information throughout the processing of input such that the current word acts as an input for the prediction of the next word. The programming language used will be Python 3 and machine learning techniques. This paper will also elaborate on the functions and structure of the various Neural networks involved.***

Keywords: CNN, LSTM, image, caption, deep learning

I. INTRODUCTION

For questions on paper guidelines, please contact us via e-mail. Image captioning works for converting a given input image into a natural language description. It is described as one of the challenging yet fundamental tasks.

This is due to it great potential impact which includes:

1) Providing compact and accurate information of images in video surveillance systems.
2) Helping in the generation of captions while sharing images on social networking sites.
3) For better understanding of content images on the web for visually impaired people.

In this paper, we analyse a deep neural network-based image caption generation method. We can provide as input the image to obtain an English/Hindi sentence describing the contents of the image. This is done through a subfield of machine learning concerned with algorithms working like the brain, structurally and functionally called Deep learning. The techniques used will be Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

CNN helps in classification based on differentiation of images from one another. The neural network consists of several convolutional layers mixed with nonlinear and pooling layers such that as an image is passed through one convolution layer, the input for the second layer is the output generated by the first layer. This is continued for all subsequent layers until we receive a fully connected layer containing output information. LSTM (RNN) can extract relevant information throughout the processing of inputs such that input of one word for each LSTM layer results in prediction of the next word, thus optimizing itself by learning from captions.

The CNN-LSTM architecture basically involves using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction such that combination of image feature and LSTM are added as inputs in the decoder model to generate output as caption with the length as that of dataset captions.

This model is specifically designed for sequence prediction with:

a) Spatial inputs, like the 2D structure or pixels in an image or the 1D structure of words in a document [4].
b) Temporal structure in their input such as the order of images in a video or words in text, or require the generation of output such as words in a textual description [4].

II. PROBLEM STATEMENT

To develop an image caption generator, an application of Deep learning requires passing the image to the model for processing and generating its description. Convolution Neural Network and Recurrent Neural Network are used to understand the content of the image and turn the understanding of the image into the work in the right order.

III.LITERATURE REVIEW

Past papers have suggested neural models, which have generated captions by using the technology of recurrent neural networks, usually a long short-term memory (LSTM). A NIC model has been used to showcase an end-to-end neural network model to not only automatically see a photo but produce a reasonable description in English using CNN which can help visually impaired people understand content of any image. This is done through a given query image to join current human-made expressions retrieved by model to produce a novel description for the inquiry picture. Further, the creation technique explains the content of the images by anticipating the nearest possible nouns, preposition and verbs.

In recent years, there has been development in the process of image description which can be seen by the [1] use of attention mechanisms to allow the neural network to have the ability to focus on its subset of inputs (specific inputs) rather than the whole.

The mechanism can be divided into two aspects:

1) Decide the part of the input to be paid attention.

2) Allocation of limited information processing resources to a significant part.

Over the time, different types of attention have been used, thus not only considering the relationship between the state and the predicted word, but also considering the image, but allowing a direct association between the title word and the image region. Therefore, recent input images have shown that the [4] trained model could detect relationships between various objects in images as well as the actions of those objects.

For evaluation,[1] BLEU and METEOR are used for machine translations, ROUGE is for automatic summary, and CIDEr and SPICE are used for image caption.

Finally, we have CNN used for extracting features from the image using a pre-trained VGG-16 model and GTTS API is used to generate the image caption to audio. Evaluation of this model is done by generating descriptions for all photos in the test dataset and evaluating those predictions with a standard cost function which functions as a standalone module.

Past papers have revealed various tools and technologies for effective image caption generation but also have some gaps to be yet filled. In paper [3], the model is only able to generate description sentences corresponding to multiple main objects for images with single target objects and also the speed of training, testing, and generating sentences for the model are not optimized. The model [3] needs to expand its scope to train on datasets larger than 100,000 images thus producing better accuracy models on production level. Right now, it is restricted to predicting the words in its vocabulary only in a single language. One more model [5] that is studied deals with cross-lingual caption generation i.e., converting English to Japanese with cross-lingual retrieval of information. This is achieved by exploiting the English corpus with establishing a connection the dataset to be a comparable corpus. When pre-training was done ahead of time, cross lingual transfer was effective when the resource in the target language and convergence was faster in same amount of time. Paper [5] serves as a baseline for future research in the area of cross-lingual image caption generation. Another paper [6] studied deals with and end to end neural network system, NIC based on CNN and RNN such that model is trained to maximize the probability of a sentence for any given image. Evaluation like [1][3][4] is done using BLEU and ranking metrics such that as dataset for image increases so does the performance. Approximately five datasets have been used to understand the effect of dataset size on generalization. One of the challenges faced involves overfitting providing high variance.

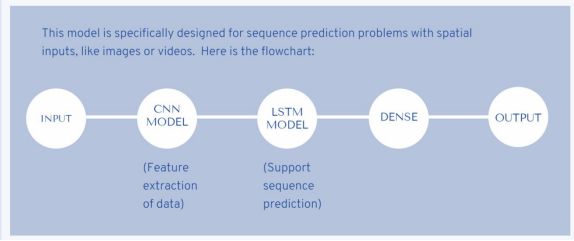


Fig. 1 Flowchart of CNN-LSTM Model

IV.DATA COLLECTION

For data, we have selected the following Kaggle dataset: https://www.kaggle.com/hsankesara/flickr-image-dataset.

Once we gain the success in creating a highly accurate model over this dataset, we wish to expand the ability of our model to process real time incoming data. This would add dynamism to our project. This is a very useful dataset with about 30k images and it has 3 attributes.

V. PROPOSED SYSTEM

Before we embark upon the actual implementation, the **5W1H** analysis is done as follows:

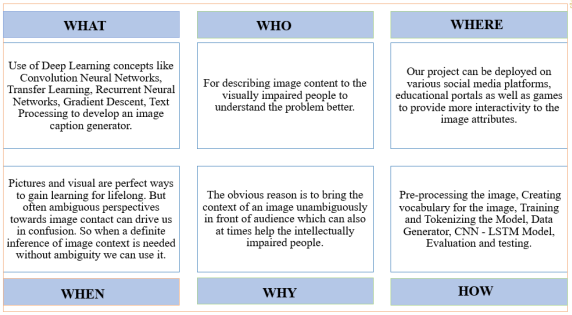


Fig. 2 5W1H Model

After a careful analysis as shown below, it can be seen that the prime requirement of clients is to express proper unambiguous meaning of images and that will perhaps help intellectually impaired people. This analysis is done after observing responses to a survey created by us and floated among college students and elders.

1) Less Critical
2) Critical
3) Recommended
4) More Critical
5) Most Critical

TABLE I
Requirement Analysis

DESCRIPTION	CRITICALITY LEVEL
Expressing image meaning	5
Removing image ambiguity	4
Suggesting good image captions	3
Aiding intellectually impaired people	4

The overview of the project can be best known by this business canvas:

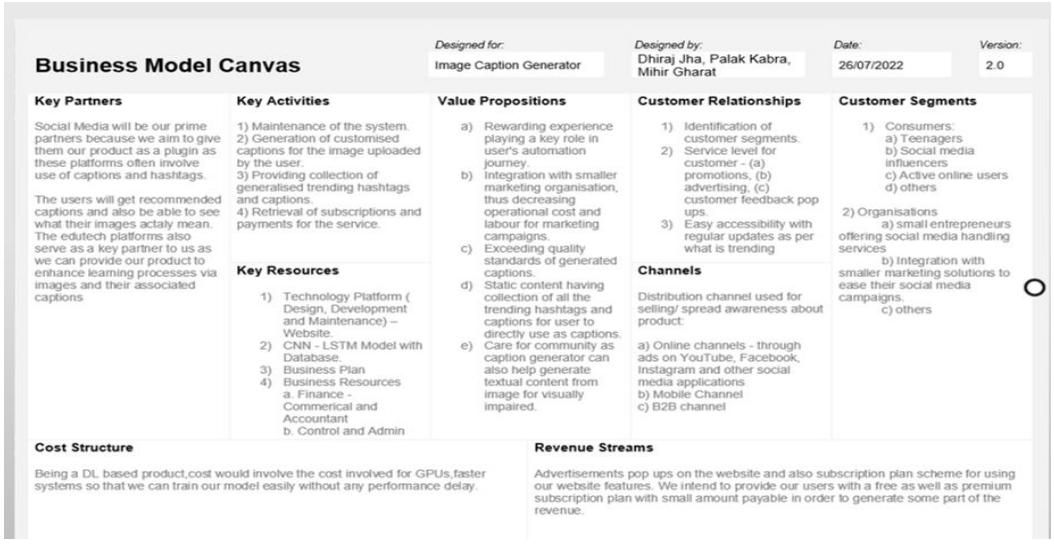


Fig. 3 Business Canvas

A. *Steps in Detail*

1) *Pre-processing of the image:* Our project would be using a pre-trained model called Visual Geometry Group (VGG16) which is pre - installed in Keras library and is used for image detection. As this is the model used to predict a classification for a photo, the features of the image are extracted just before the last layer of classification.
2) *Creating of Vocabulary for the image:* Cleaning of dataset is a pre-requisite and is performed by splitting it into words for easy handling of punctuation and case sensitivity issues. As computers use binary language and are yet not competent to use English words, we have to represent them such that each word of the vocabulary is mapped onto a unique index value followed by encoding each and every word into a fixed size vector and representing each word as a numerical value. This will ensure that the text is readable by the machine and then eventually generate the captions for the image. Data cleaning can be done by:

a) Loading the data.
b) Creating a descriptions dictionary that maps images.
c) Removing punctuations, converting all text to lowercase and removing words that contain numbers.
d) Separating all the unique words and creating vocabulary from all the descriptions.
e) Creating a descriptions.txt file to store all the captions.

3) *Training the Model*
4) *Tokenizing Model:* Keras provides the Tokenizer class that can learn this mapping from the loaded description data. This will fit a tokenizer given the loaded photo description text. We need to map each word of vocabulary with a unique index value. Keras library provides a function that will be used to create tokens from vocabulary and then to a tokenizer.pkl pickle file.
5) *Data Generator:* As this is a supervised learning task, providing an input and output to the model is required for training. We train our model on nearly a hundred thousand images such that each image contains:

a) 4096-length feature vector.
b) Corresponding caption for the image is represented in the form of numerical values.

6) *CNN - LSTM Model:* To train the model, we will be using a hundred thousand training images, thus generating its input and output sequences in batches followed by fitting them to the model. 10 epochs would be considerable to train the model. The features vectors obtained from the VGG Network will aid in developing an LSTM based model which will be working towards getting the desired outcome - sequence of words for the given image called caption.
7) *Evaluation of Model*
8) *Testing the Model:* After training of model is completed, testing of model against random images is required using the same tokenizer.pkl file. The predictions contain the maximum length of index values, thus the same tokenizer.pkl will aid in getting the words from their index values.
9) *Deployment and Front-End*

VI. RESULT AND DISCUSSION

The method's end-to-end learning structure is its strength. The flaw is that it necessitates a big amount of human-labelled data, which is too expensive in practice. Furthermore, both object detection and phrase production are still subject to significant errors using the existing technique. The text and image files are loaded into distinct variables in our software, while the test file is saved in a string. This string is used and altered to generate a dictionary that associates each image with a set of five descriptions.

Here are few screenshots of the developed front end which is a website developed using HTML CSS and JS.

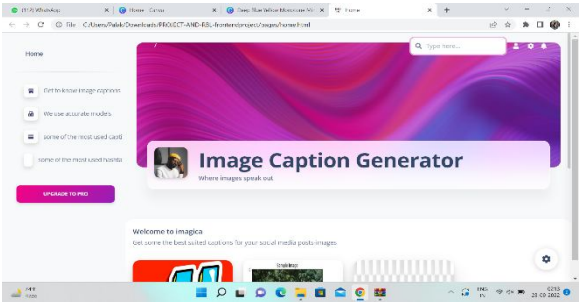


Fig. 4 Home Page

This is our homepage. A navigational menu to various sections is shown i.e., Hashtags, Caption Generator and Caption Recommender.

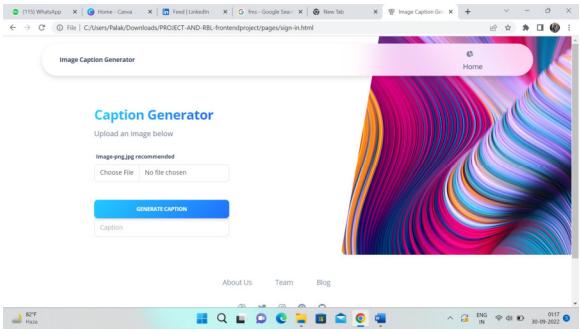


Fig. 5 Image Caption Generator Page

Here, we give users an interface where they can upload an image and then after hitting the Generate Caption button, get the caption for the image. This page will be linked to our Model (CNN+LSTM) created and integrated via flask.

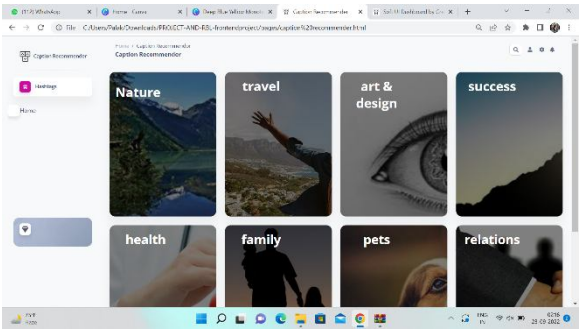


Fig. 6 Static Caption Recommender Page

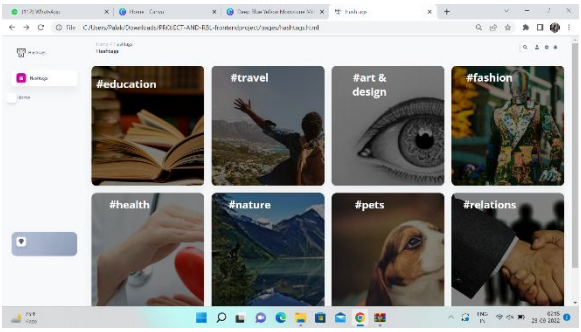


Fig. 7 Static Hashtag Recommender Page

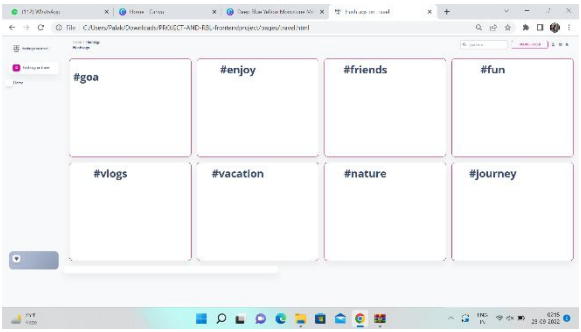


Fig. 8 Hashtag Page

Here we can see that once a user clicks on hashtags, they see some domains of which they can view some widely used hashtags like Nature, travel, etc. The same is done for Captions.

Currently these pages just give a static list of hashtags and captions but in future we wish to make it dynamic to show captions and hashtags that are trending based on real time scenarios. The UI template has been referenced from [7] creativetimofficial GitHub repository.

VII. CONCLUSIONS

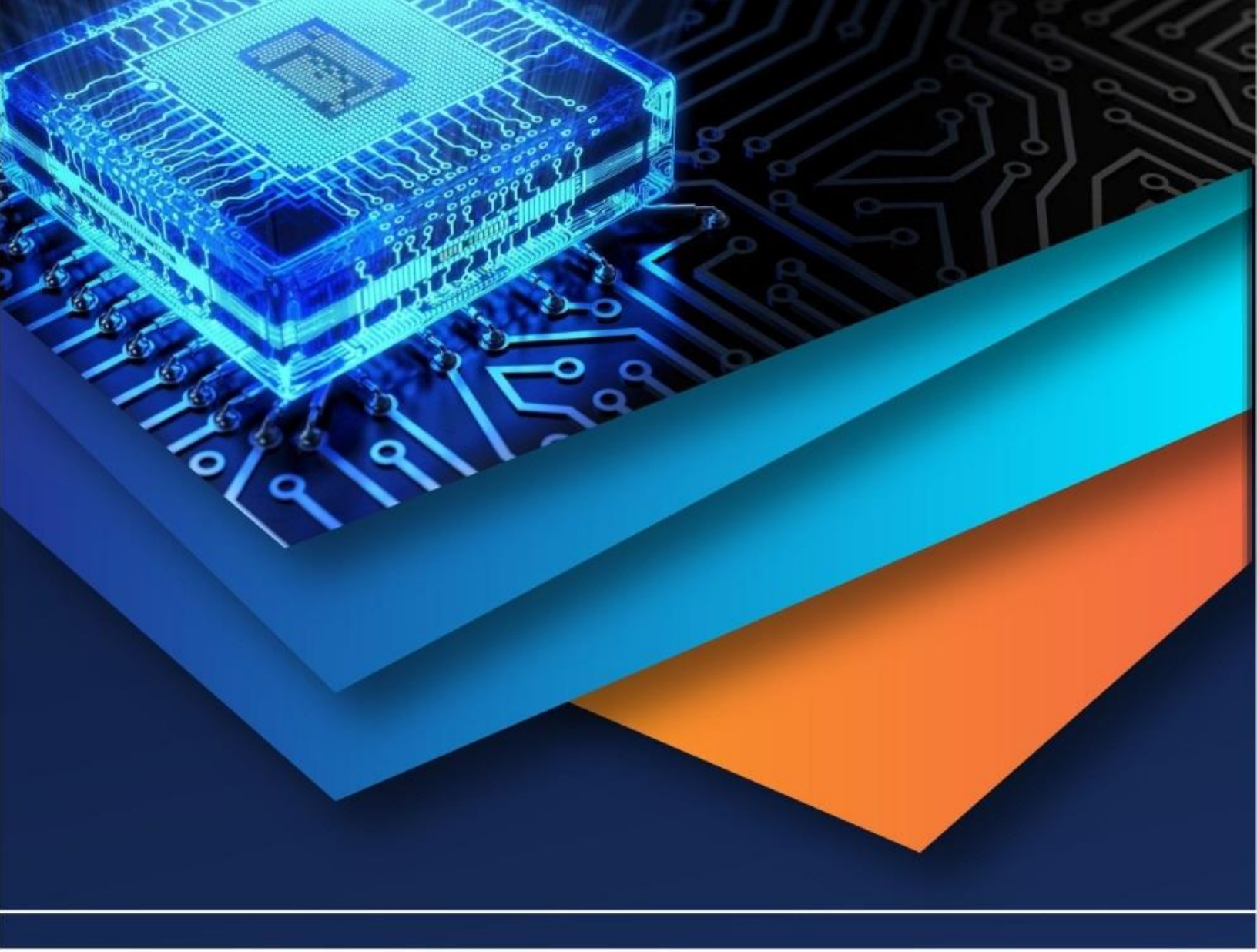
As a result, we can conclude that deep learning can be used to generate image captions. We can go even further by creating a hashtag generator. Based on the findings, we may conclude that the deep learning technology employed yielded positive outcomes. Because the CNN and the LSTM were synchronized, they were able to determine the relationship between objects in images.

VIII. ACKNOWLEDGMENT

We gratefully acknowledge the support, guidance and encouragement of our mentor Mr. Shailesh Sangle for this work.

REFERENCES

[1] Haoran Wang ,Yue Zhang and Xiaosheng Yu, An Overview of Image Caption Generation Methods, Published: 09 Jan 2020
[2] http://sersc.org/journals/index.php/IJAST/article/view/5927/3650
[3] Krishnakumar , K.Kousalya, S.Gokul, R.Karthikeyan D.Kaviyarasu , Image Caption Generator using Deep Learning, International Journal of Advanced Science and Technology, Vol. 29, No. 3s, (2020), pp. 975-980ISSN: 2005-4238 IJAST
[4] P. Aishwarya Naidu1, Satvik Vats,Gehna Anand, Nalina V, A Deep Learning Model for Image Caption Generation, Published: 30/June/2020
[5] Takashi Miyazaki, Nobuyuki Shimizu, Yahoo Japan Corporation Tokyo, Japan, Cross- Lingual Image Caption Generation
[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan , Show and Tell: A Neural Image Caption Generator. CVPR2015
[7] https://github.com/creativetimofficial/soft-ui-dashboard



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY