B.Tech Thesis Project Mid Term Presentation

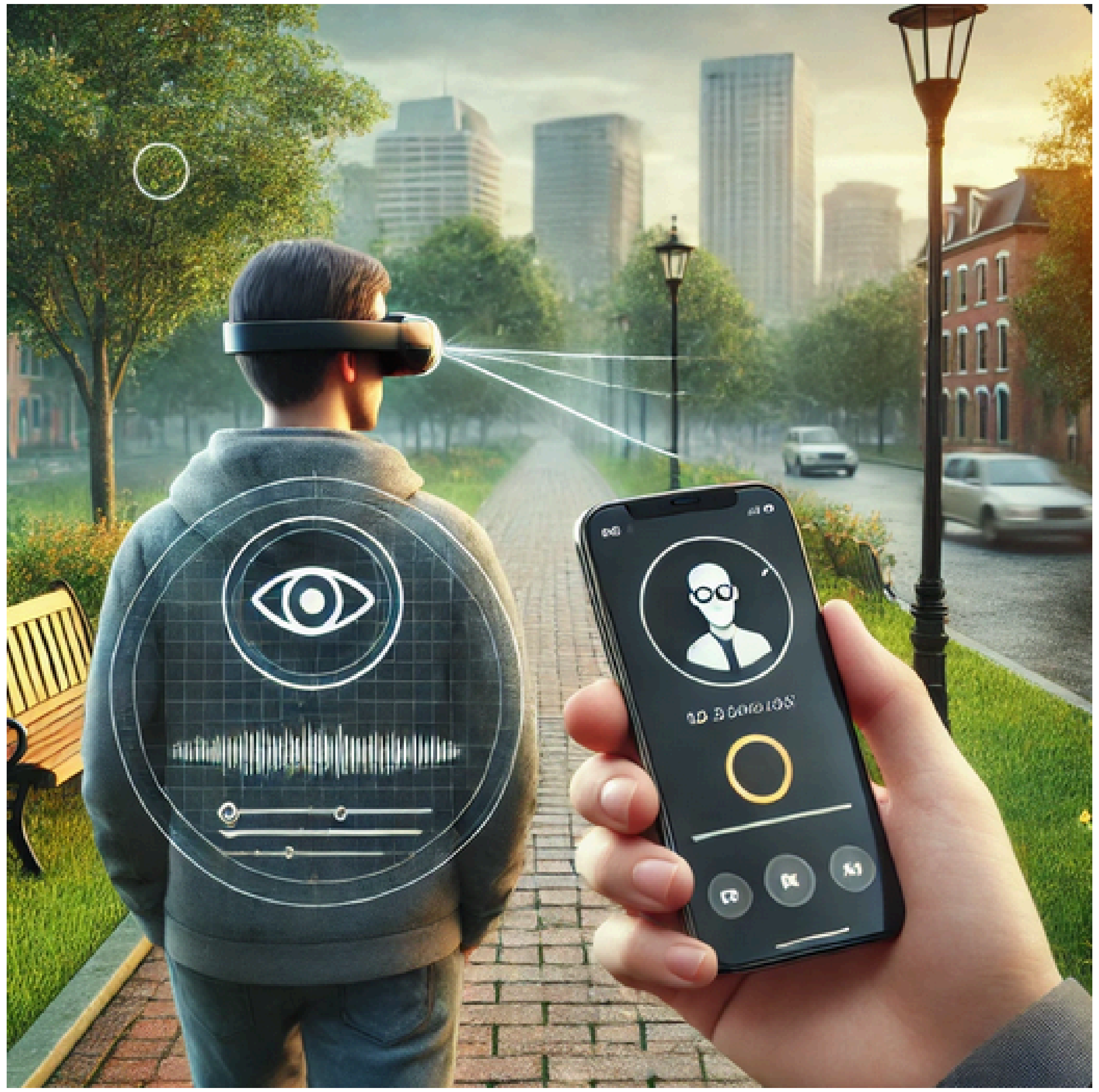
Image Description Generator for Assisting Visually Impaired Individual

Submitted By:
Muskan Debnath
2021BCS-046

Supervisor:
Prof. Shashikala Tapaswi

INTRODUCTION

The "Image Description Generator for Assisting Visually Impaired Individuals" aims to leverage AI-driven computer vision and NLP to generate accurate, context-aware image descriptions. Using an encoder-decoder neural network with attention mechanisms, the system enhances accessibility by providing real-time image-to-audio conversion via a mobile app. With over 2.2 billion visually impaired people globally (WHO), this project addresses a critical need for inclusivity while optimizing model efficiency and usability.



PROBLEM STATEMENT

- To study and develop a solution for predicting the description of the images in real-time with good accuracy.
- To further modify the existing models on image description generation with new modifications to reduce the complexity and training time of the proposed model.
- To develop a native application for fetching the image description given the image.

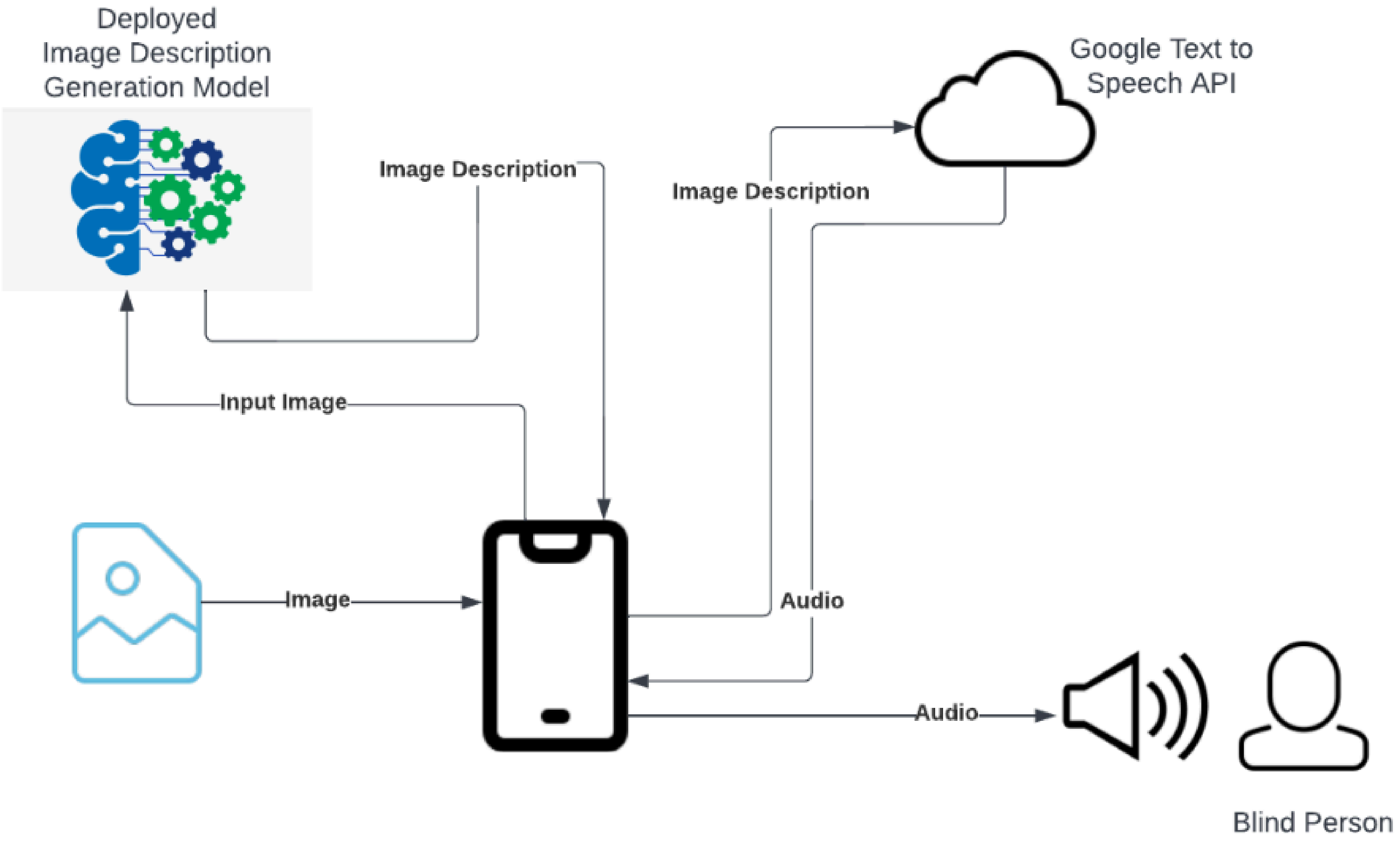


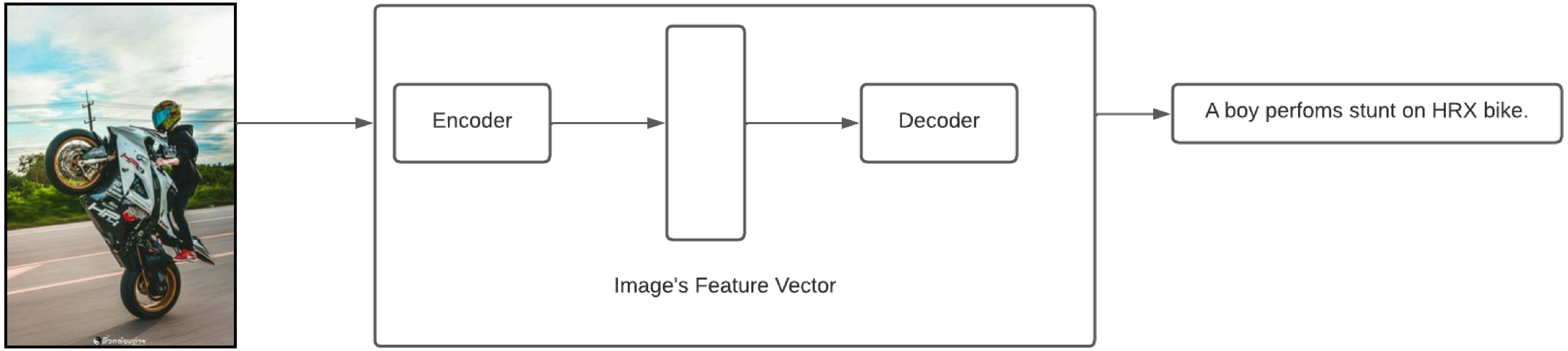
Figure: An overview of the proposed architecture of the image description generation module.

SYSTEM OVERVIEW & METHODOLOGY

- The system follows a Neural Network-based Encoder-Decoder architecture.
- Main Components:
    - Feature Extraction (Encoder):
        - A Convolutional Neural Network (CNN) extracts feature representations from images.
        - Pre-trained models like VGG16, Inception V3, and ResNet50 can be used.
    - Description Generation (Decoder):
        - A Recurrent Neural Network (RNN) generates meaningful captions.
        - LSTM (Long Short-Term Memory) are tested.
    - Attention Mechanism:
        - Ensures the model focuses on important image regions while generating captions.
        - Helps in making descriptions more contextually relevant.

FEATURE EXTRACTION

Most of the studies conducted on image captioning have proved that encoder-decoder based architecture gives a rich description of the image with high performance. The three dimensional image is given as input to the encoder part of the architecture and image’s feature representation is extracted as output of the encoder. Convolutional Neural Network (CNN) serves the purpose of the encoder. Generally, CNN performs manipulations on the image using the Convolution layer and Pooling Layer. It effectively reduces the dimensionality of the image without losing the necessary information or features captured in the image.



Encoder

Image's Feature Vector

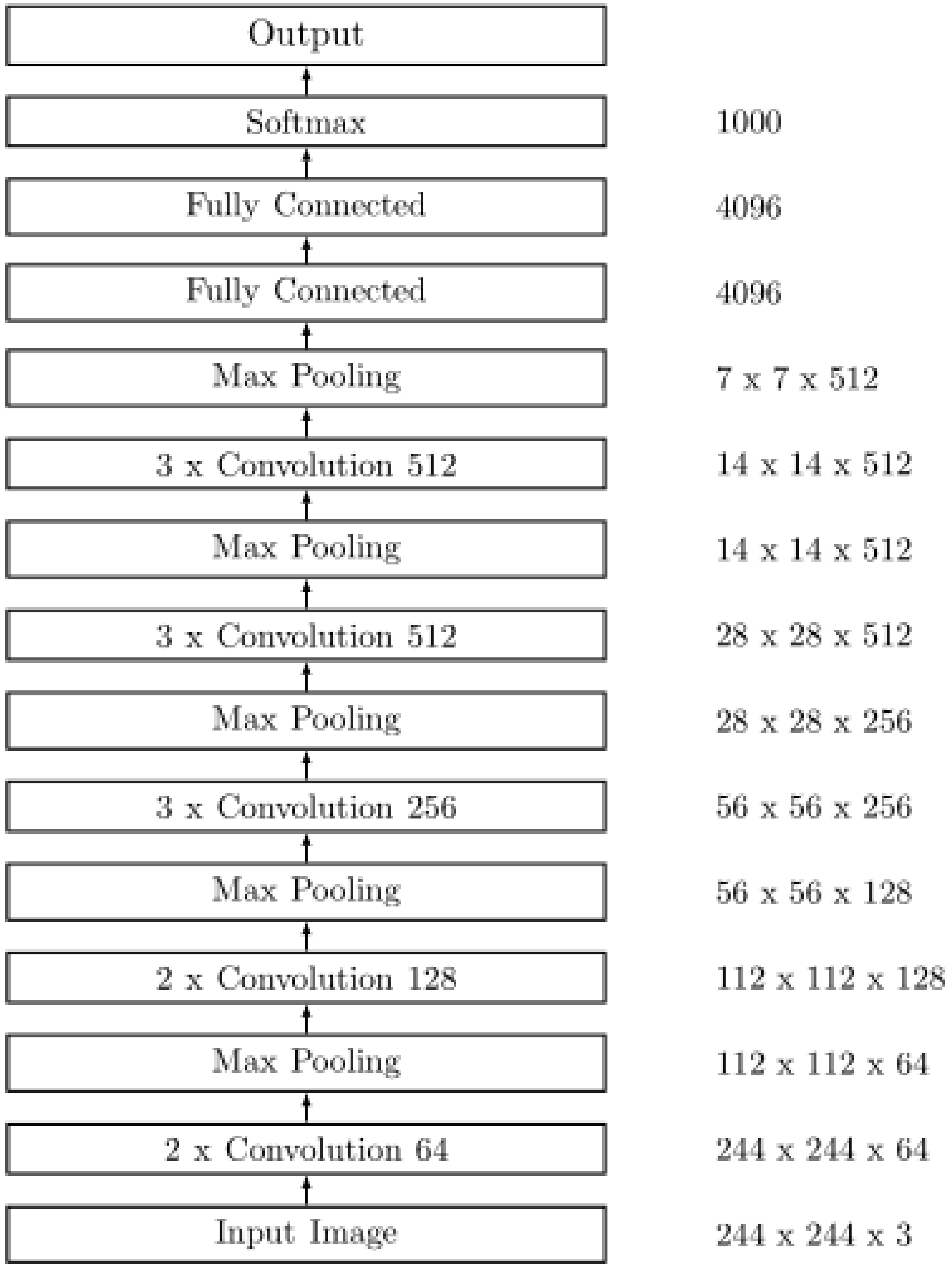
Decoder

A boy perfoms stunt on HRX bike.

VGG 16

Visual Geometry Group (VGG) is a Convolutional Neural Network (CNN) architecture proposed by Karen et al. VGG16 has 16 number of layers where weights are learned while training the model.

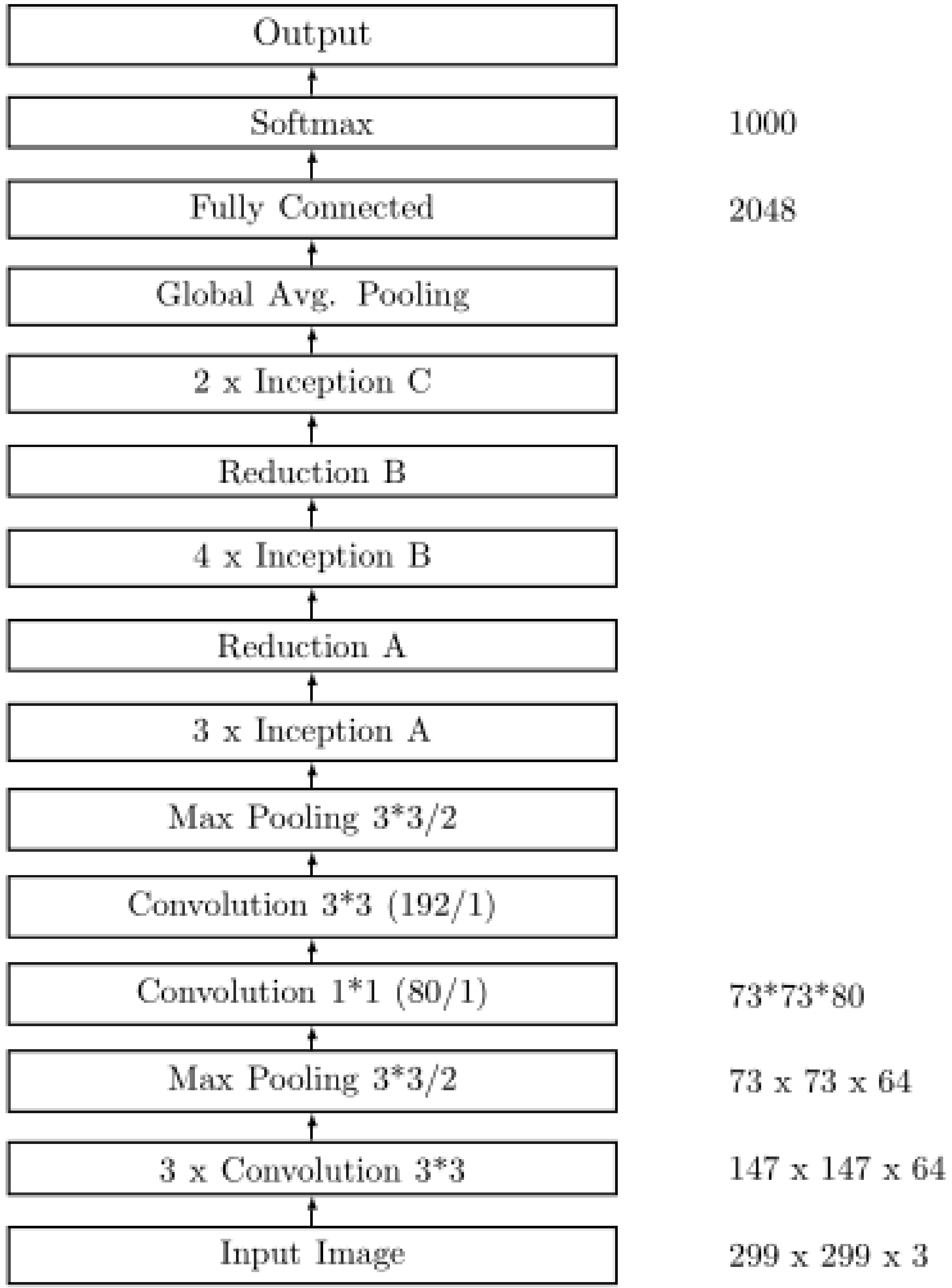
The VGG16 model takes a standard input image with size 244x244x3. **It uses kernel filter of 3x3 size with stride 1. They use maxpool of size 2x2 with stride 2. The size of the kernel filter, maxpool filter and padding is constant throughout the model.** The detailed architecture of VGG16 is given in the figure



Output	
Softmax	1000
Fully Connected	4096
Fully Connected	4096
Max Pooling	7 x 7 x 512
3 x Convolution 512	14 x 14 x 512
Max Pooling	14 x 14 x 512
3 x Convolution 512	28 x 28 x 512
Max Pooling	28 x 28 x 256
3 x Convolution 256	56 x 56 x 256
Max Pooling	56 x 56 x 128
2 x Convolution 128	112 x 112 x 128
Max Pooling	112 x 112 x 64
2 x Convolution 64	244 x 244 x 64
Input Image	244 x 244 x 3

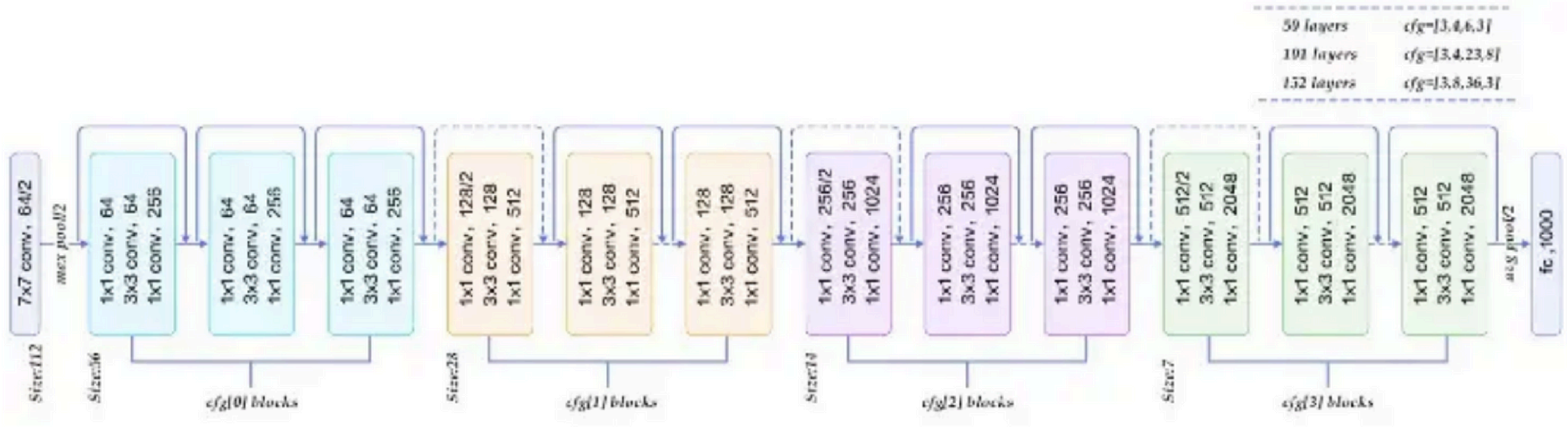
INCEPTION

Inception V3 is also a CNN pre-trained model published in 2015 [3]. It performed better than its subsequent parts like V1 and V2. There were many major advancements that led to increase in the performance of Inception V3. It factorized the convolutions into many smaller convolutions, convolution was spillted to many asymmetric convolution to reduce the computation. **In total, the Inception V3 consisted of ~48 layers.**



RESNET50

The ResNet architecture was proposed to solve the problem of the vanishing gradient commonly seen in the CNN models. The problem was resolved using the skip connections between the different layers . ResNet50 architecture is 50 layers deep and had fewer parameters than the VGG16 and Inception V3 architecture. Though, the model will be difficult to train because of the increase in the depth but will certainly have better performance.



FEATURE VECTOR REPRESENTATION OF IMAGE

Consider, image I is fed into any selected convolutional neural network model. Let’s say F is the feature vector received as output after fully connected layer of CNN and V is the spatial vector received as output from the last convolution layer of CNN.

The output of the last convolution layer V is the visual k x k grid which can map to the previous convolution layer outputs and can perfectly define the relative feature positions.

The fully connected layer output F will be given as input to the decoder or language model in model where no attention mechanism is adopted.

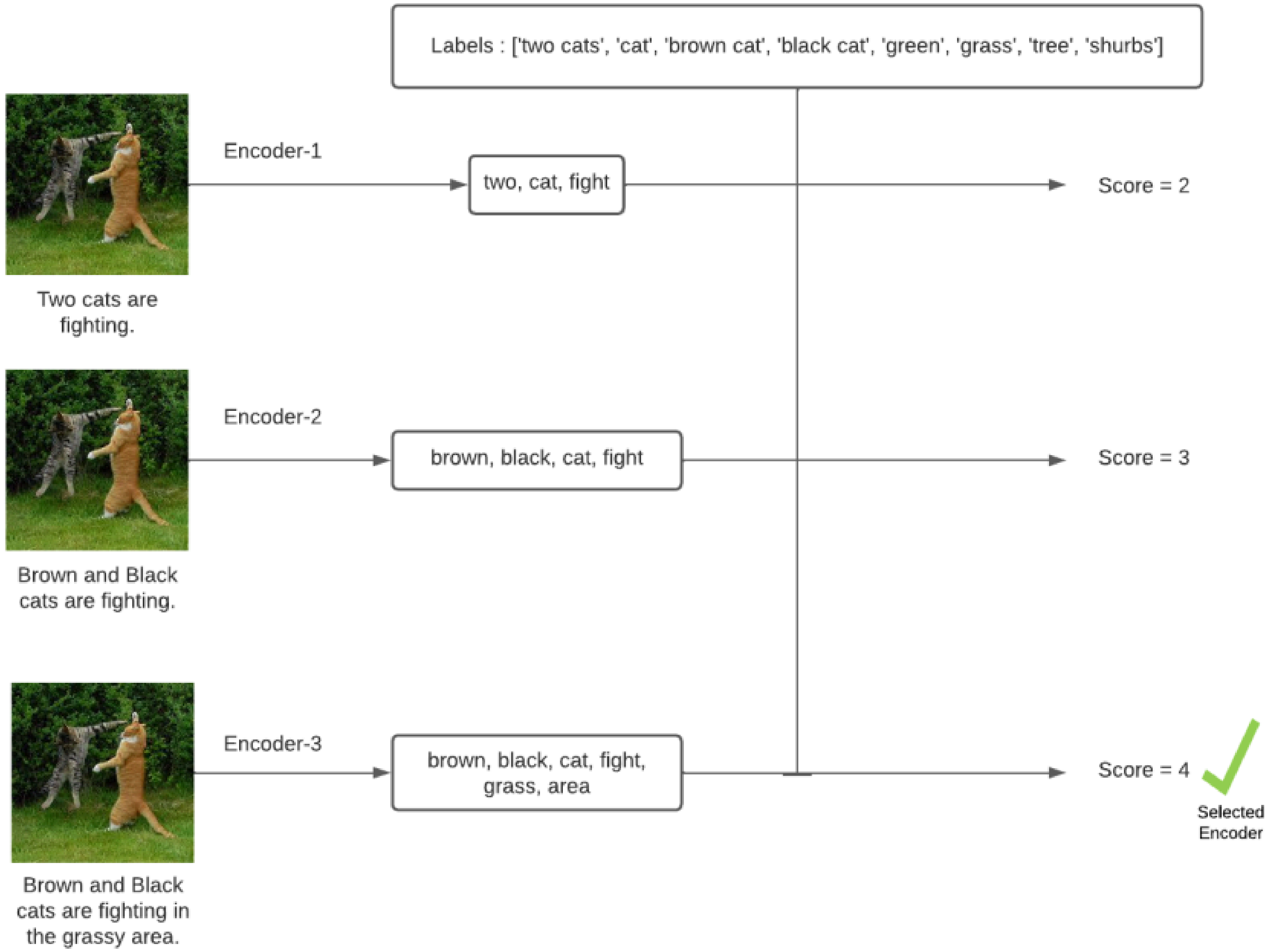
$$F = C_{fc}(I)$$

$$V = C_{conv}(I)$$

$$V = \{v_1, v_2, ..., v_{k^2}\}$$

AUTOMATING ENCODER SELECTION

There are various encoders that can be used to fetch the image vector representation. To deal with the uncertainty of selecting the encoder that gives the best result for image description generation, we deploy all models with different encoders and automate the selection based on matching labels obtained from the YOLO object detection algorithm. Labels are generated by YOLO, and captions are generated by various encoder-decoder models. Root words are extracted using NLTK, and a score is calculated based on the number of root word matches with the labels. The model with the best score is selected for sending the image description.



Labels : ['two cats', 'cat', 'brown cat', 'black cat', 'green', 'grass', 'tree', 'shurbs']

Encoder-1

two, cat, fight

Score = 2

Two cats are fighting.

Encoder-2

brown, black, cat, fight

Score = 3

Brown and Black cats are fighting.

Encoder-3

brown, black, cat, fight, grass, area

Score = 4

Selected Encoder

Brown and Black cats are fighting in the grassy area.

DESCRIPTION GENERATION

The task of image description generation deals with the generation of sequence of words. The idea of generation of sequence of words make us use the Recurrent Neural Network (RNN). The naive RNN network fails to deal with the long sequence of words because of the exploding gradients words get vanished with time. Many researches conducted on image caption generation adopted LSTM as the language generation model as it solves the problem discussed aforementioned.

Simple RNNs struggle with long-term dependencies due to the vanishing gradient problem, making them ineffective for long sequences. LSTM overcomes this by using cell states and gating mechanisms (input, forget, output gates) to retain important information over longer sequences.

DESCRIPTION GENERATION

Long Short Term Memory - LSTM

- It consist of three gates: input gate, output gate and the forget gate.
- The value of previous hidden state, ht−1 and the current element of the sentence xt is passed through the sigmoid function.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

- Now, we need to decide through the input layer, which values needs to be updated anset of all the possible candidates that could be the next potential Ct .

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C.[h_{t-1}, x_t] + b_c)$$

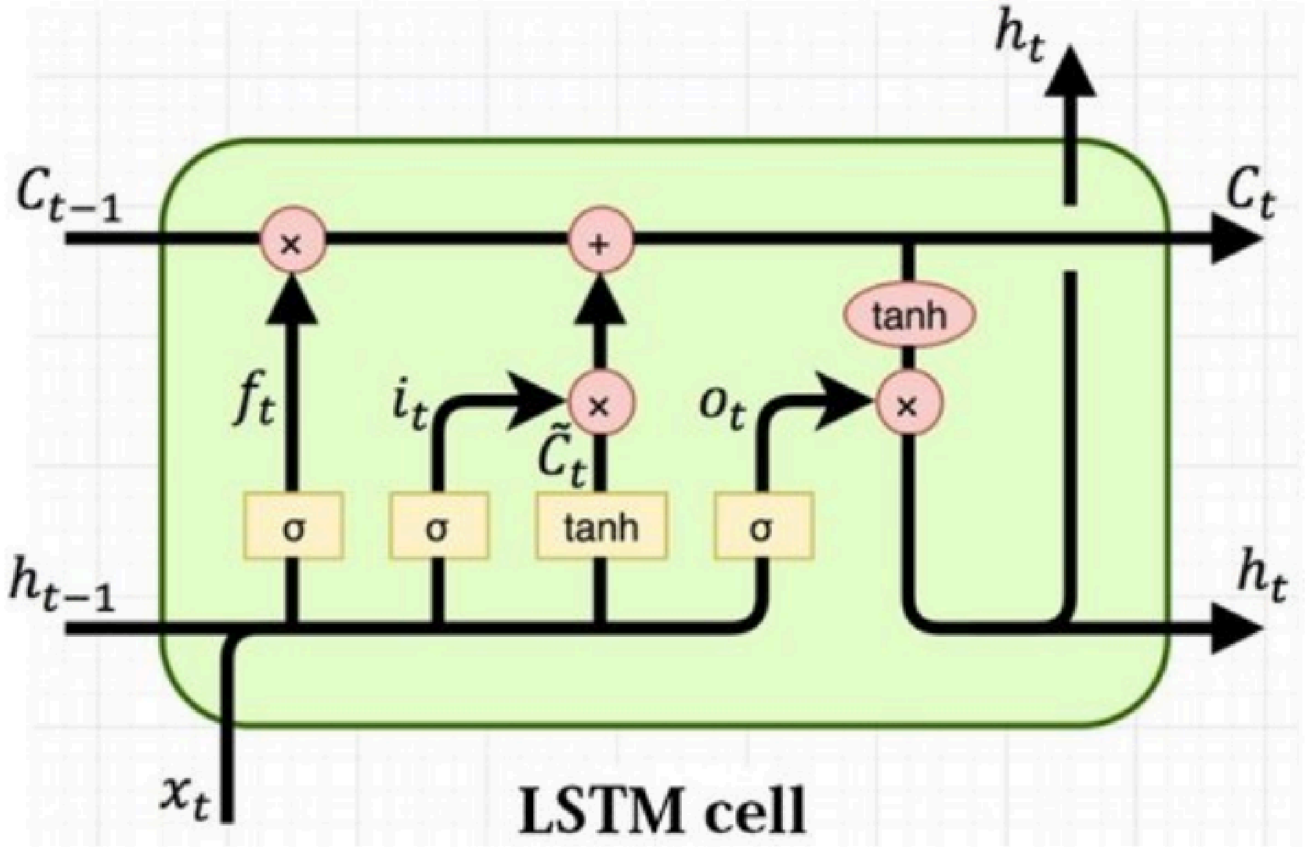


Figure: The basic architecture of the repeating module of LSTM [1]

DESCRIPTION GENERATION

Long Short Term Memory - LSTM

- Ct−1 will be updated to Ct using the formula given below.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- The output of the hidden state of this iteration will be decided using the sigmoid and the hyperbolic function.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

The Ct and ht will be fed as input to the LSTM at t + 1 iteration. At every iteration, the ht given to the softmax function to get the distribution of probability of every word from the dataset.

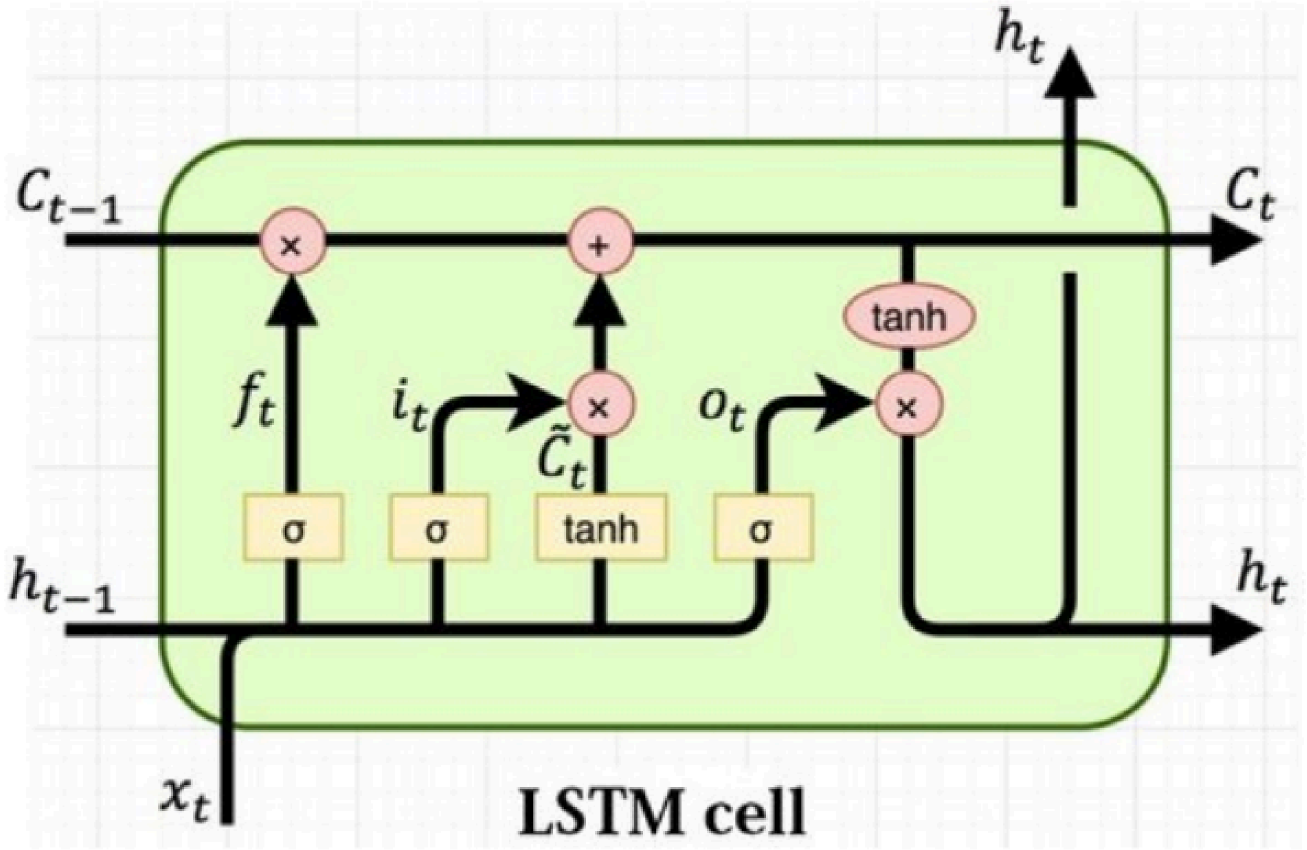


Figure: The basic architecture of the repeating module of LSTM [1]

VISUAL ATTENTION

- In attention-based image description generation, the output of the convolutional fully connected layer fed into the attention module.
- With attention gate, the decoder concentrate on only those pixel of the image that are relevant.
- In each iteration, the feature vector representation of the image and previous context of the LSTM is given to the attention module.
- The training of the attention mechanism can be done using the backpropagation.

$$a_t = \sum_{1}^{n} S_j Y_j$$

ABOUT FLICKR 8K DATASET

- Description: A collection of 8,000 real-world images from Flickr, covering diverse scenes and activities.
- Captions: Each image has five human-generated descriptions, ensuring linguistic diversity.
- Purpose: Used for image captioning, vision-language modeling, and automatic image description generation.
- Applications:
  - Helps train AI models for image-to-text conversion.
  - Useful in assistive technologies for visually impaired individuals.
  - Supports content-based image retrieval and AI-driven accessibility.
- Format: Available in CSV or JSON, making it easy to use in deep learning models.
- Challenges: Requires deep learning architectures for meaningful caption generation and context understanding.

FUTURE WORK

- **Fine-Tuning the Model with Domain-Specific Data:** To further enhance accuracy, the model will be fine-tuned using domain-specific datasets tailored to real-world applications. This approach will help the model generate more context-aware and precise captions, improving its performance across different scenarios.
- **Training the Model on Large-Scale Datasets:** The final model will be trained on extensive datasets like Flickr30k and MS-COCO to enhance its accuracy and generalization. These datasets provide diverse image-caption pairs, helping the model generate more precise and meaningful descriptions.

INTERNSHIP DETAILS

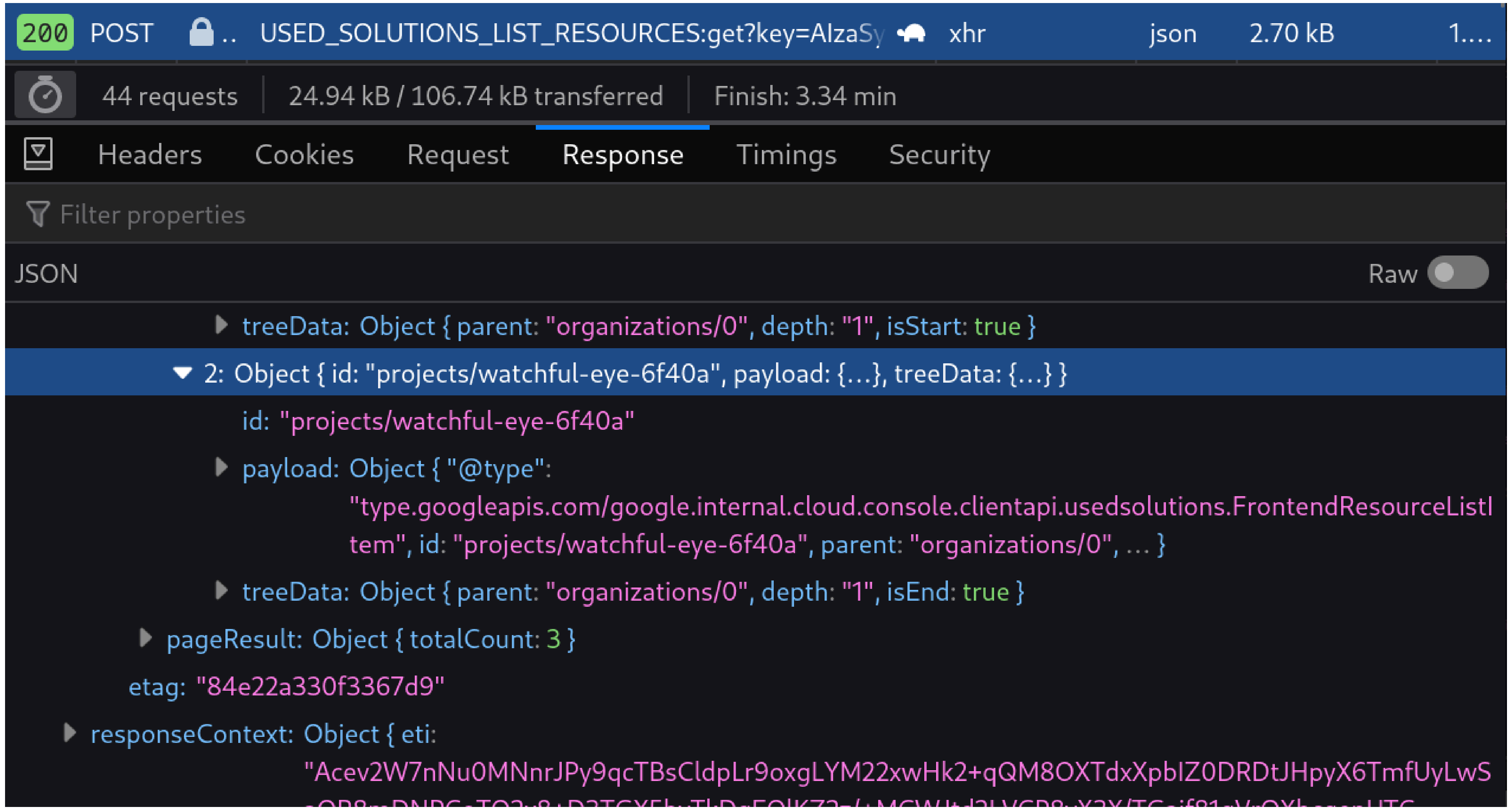
zscaler™

COMPANY OVERVIEW

- Zscaler is a cloud-based security company that provides secure internet and web gateway services for businesses and organizations.
- The company's flagship product, Zscaler Internet Access (ZIA), enables secure access to the internet and cloud applications for users anywhere, on any device.
- Additionally, Zscaler offers solutions for secure access to private applications (ZPA) and workloads, as well as advanced threat protection against cyber threats and data breaches.
- Zscaler operates on a Zero Trust architecture, ensuring least-privileged access, eliminating attack surfaces, and preventing lateral movement within networks.

WORK DETAILS

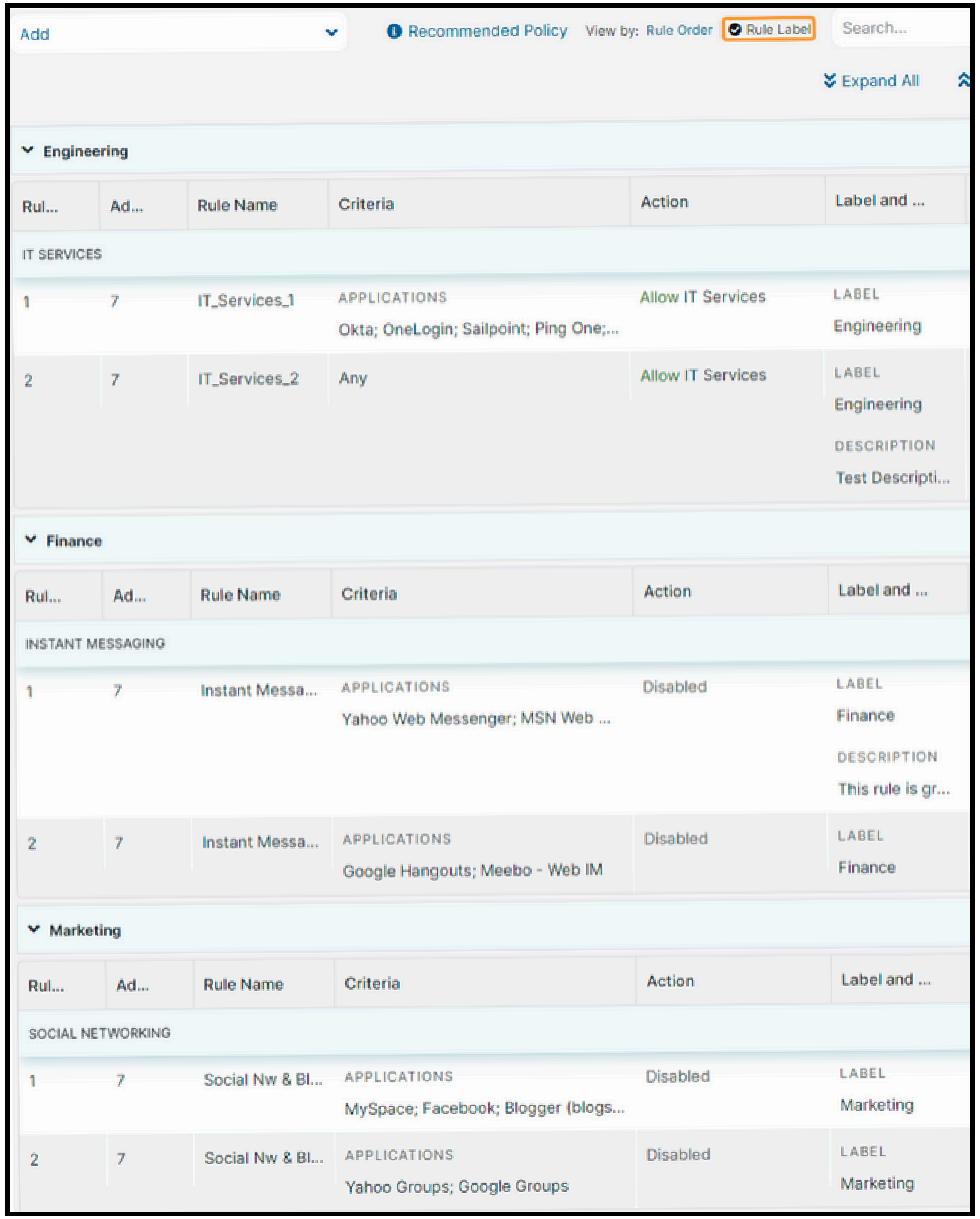
Parsing Headers, Cookies, and Transactions for Multi-Level Instance Detection: Developed methods to extract multi-level instance values by parsing headers, cookies, and transactions in HTTP requests and responses. This process improved real-time instance discovery and policy enforcement, ensuring better control over cloud resources. Enhanced cloud visibility and anomaly detection mechanisms, allowing for more accurate identification of cloud applications and their usage patterns.



```
200  POST  .. USED_SOLUTIONS_LIST_RESOURCES:get?key=AIzaSy   xhr   json   2.70 kB   1....
44 requests | 24.94 kB / 106.74 kB transferred | Finish: 3.34 min
Headers   Cookies   Request   Response   Timings   Security
Filter properties
JSON                                                                   Raw
            ▶ treeData: Object { parent: "organizations/0", depth: "1", isStart: true }
          ▼ 2: Object { id: "projects/watchful-eye-6f40a", payload: {…}, treeData: {…} }
              id: "projects/watchful-eye-6f40a"
            ▶ payload: Object { "@type":
                    "type.googleapis.com/google.internal.cloud.console.clientapi.usedsolutions.FrontendResourceListI
                    tem", id: "projects/watchful-eye-6f40a", parent: "organizations/0", … }
            ▶ treeData: Object { parent: "organizations/0", depth: "1", isEnd: true }
        ▶ pageResult: Object { totalCount: 3 }
          etag: "84e22a330f3367d9"
    ▶ responseContext: Object { eti:
                    "Acev2W7nNu0MNnrJPy9qcTBsCldpLr9oxgLYM22xwHk2+qQM8OXTdxXpbIZ0DRDtJHpyX6TmfUyLwS
```

WORK DETAILS

Implementation of Cloud Security Policies through URL Filtering: Worked on URL filtering to regulate access to cloud applications. Implemented security policies to restrict unauthorized access, ensuring compliance with organizational guidelines. Additionally, contributed to instance discovery and control, identifying cloud instances across multiple environments and enforcing policies to enhance cloud security. These efforts helped in mitigating risks associated with shadow IT and unauthorized cloud usage.



Engineering

Rul...	Ad...	Rule Name	Criteria	Action	Label and ...
IT SERVICES					
1	7	IT_Services_1	APPLICATIONS Okta; OneLogin; Sailpoint; Ping One;...	Allow IT Services	LABEL Engineering
2	7	IT_Services_2	Any	Allow IT Services	LABEL Engineering DESCRIPTION Test Descripti...

Finance

Rul...	Ad...	Rule Name	Criteria	Action	Label and ...
INSTANT MESSAGING					
1	7	Instant Messa...	APPLICATIONS Yahoo Web Messenger; MSN Web ...	Disabled	LABEL Finance DESCRIPTION This rule is gr...
2	7	Instant Messa...	APPLICATIONS Google Hangouts; Meebo - Web IM	Disabled	LABEL Finance

Marketing

Rul...	Ad...	Rule Name	Criteria	Action	Label and ...
SOCIAL NETWORKING					
1	7	Social Nw & Bl...	APPLICATIONS MySpace; Facebook; Blogger (blogs...	Disabled	LABEL Marketing
2	7	Social Nw & Bl...	APPLICATIONS Yahoo Groups; Google Groups	Disabled	LABEL Marketing

THANK YOU