IMAGE CAPTIONING

Priya Chahal

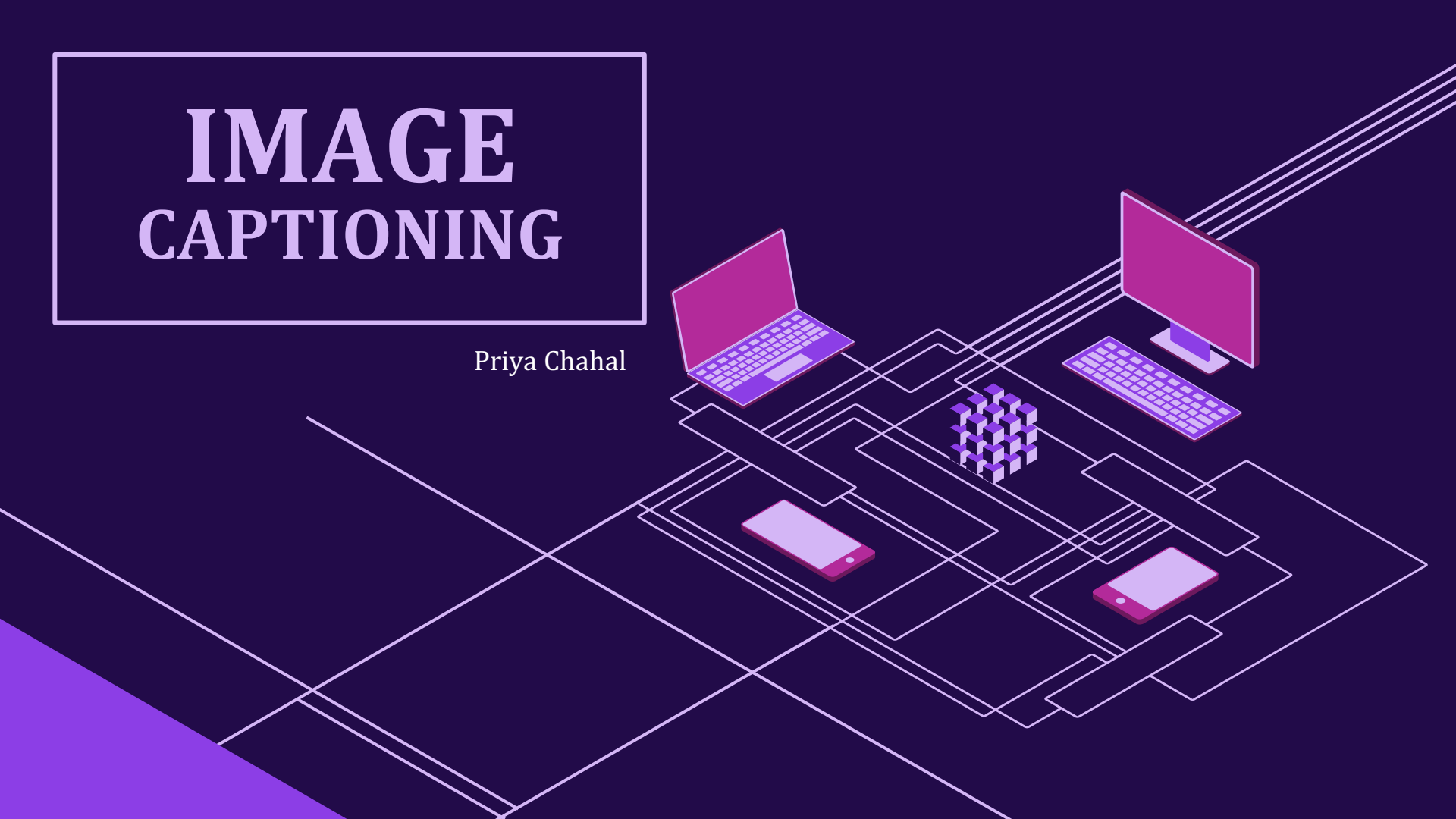


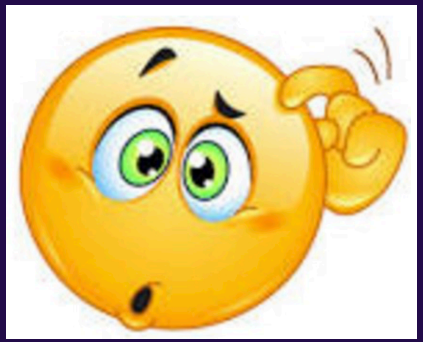
TABLE OF CONTENTS

Can you write a caption?

Humans

- **A white dog in a grassy area.**
- **White dog with brown spots.**
- **A dog on grass and some pink flowers.**

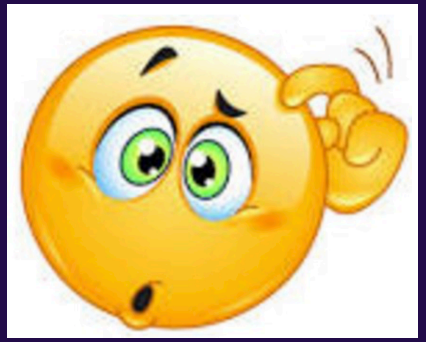
Computers



01

PROBLEM / BUSINESS STATEMENT

Problem Statement



How do we make computers understand what the image is about?

SOLUTION

IMAGE CAPTIONING

Train the computers to understand what the image is about?

But why do we want to do it??

Real World Applications

- **CCTV cameras**
  Generate alarms based on video's captions, which can help to reduce crime/accidents.

- **Healthcare sector**
  Added aid to monitor a patient movements.
  Read the X-rays, other images.

- **Self driving Cars**
  Generating captions of surroundings, can help in better self driving systems.

- **Aid to the blind**
  A product to convert "scene to text (Image captioning)" and then "text to audio".

- **Improve Google Image search experience**
  Convert Image to captions.

02

Introduction to Data & Model Flow

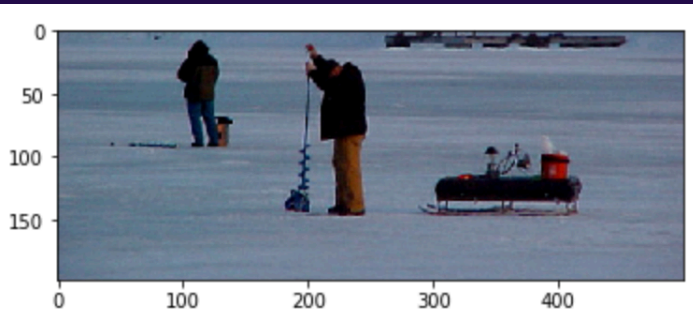
Data

Flickr 8K dataset

- 8000 images in total

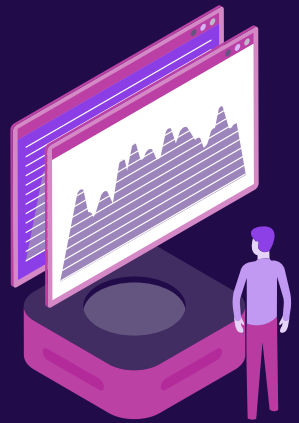
6000 train images
1000 development images
1000 test images

- 5 captions per Image



102351840_323e3de834.jpg

A man drilling a hole in the ice .
A man is drilling through the frozen ice of a pond .
A person in the snow drilling a hole in the ice .
A person standing on a frozen lake .
Two men are ice fishing .



Input

- Images
- Captions (Text)

Model

- Image features (Array)
- Text (Integers)

Model trained to predict the next word

Output

Generate captions for images.

- Greedy search
- BEAM search

03

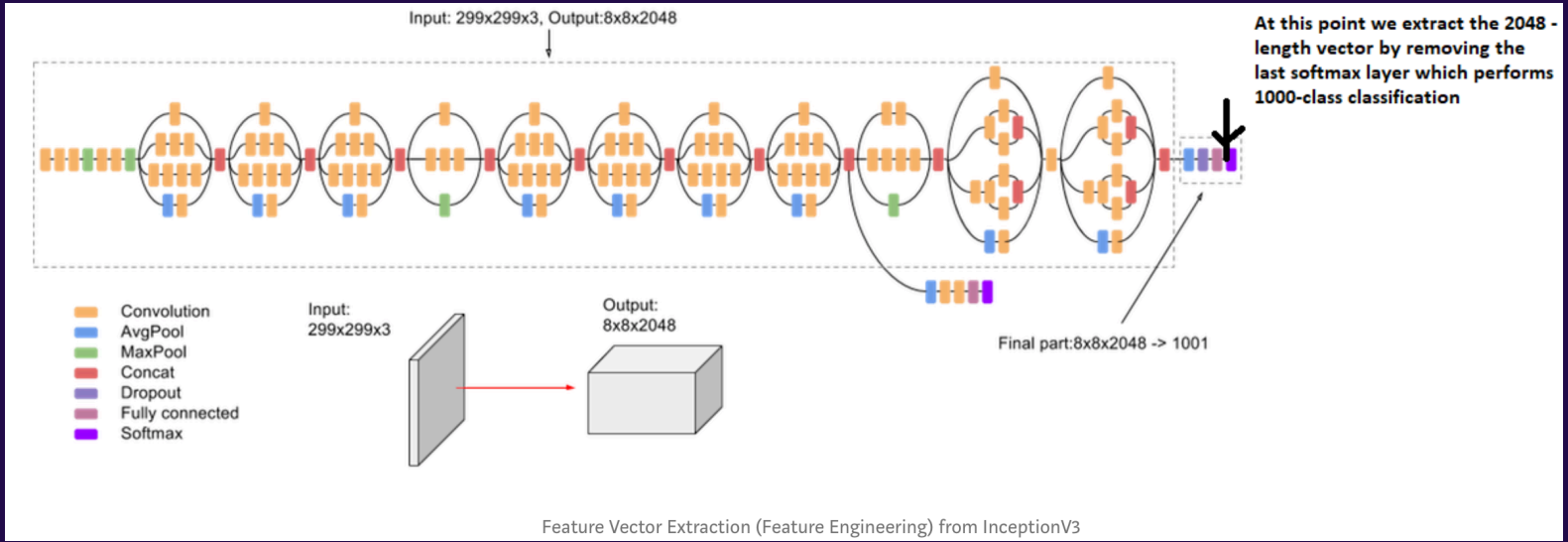
Data Pre-processing

Data Pre-processing : Feature extraction

Pre-trained Model : Inceptionv3

- CNN Model
- Developed by Google Research
- 48 layers
- classify images into 1000 object categories

Input size : (299, 299, 3)
Output size : (1, 2048)



Feature Vector Extraction (Feature Engineering) from InceptionV3

Data Pre-processing : Text to Sequence

Data Cleaning

- Lower case
- Removed punctuations
- Removed numbers, eg: 1950.

- Addition of ”startseq” and “endseq” keywords to each caption.

Caption1 : startseq black dog is running after white dog in the snow endseq

Text	startseq	black	dog	is	running	after	white	dog	in	the	snow	endseq
text_to_sequence	1	14	8	6	31	252	13	8	3	4	41	2

Input_sequence	Output sequence
'startseq' 1	black 14
'startseq' 1 'black' 14	dog 8
'startseq' 1 'black' 14 'dog' 8	is 6
'startseq' 1 'black' 14 'dog' 8 'is' 6	running 31
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31	after 252
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252	white 13
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252 'white' 13	dog 8
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252 'white' 13 'dog' 8	in 3
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252 'white' 13 'dog' 8 'in' 3	the 4
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252 'white' 13 'dog' 8 'in' 3 'the' 4	snow 41
'startseq' 1 'black' 14 'dog' 8 'is' 6 'running' 31 'after' 252 'white' 13 'dog' 8 'in' 3 'the' 4 'snow' 41	endseq 2

04

Data Modelling

Data Modelling : Train the Model(RNN-LSTM)

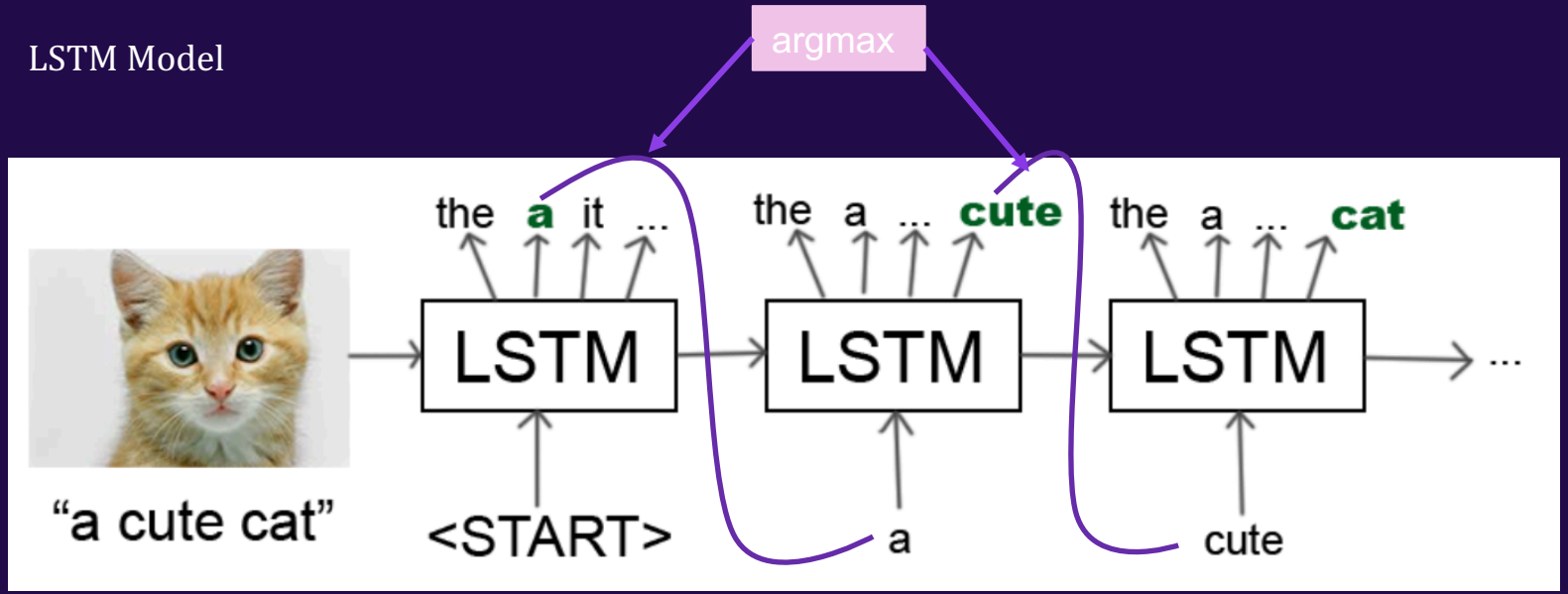
Captions → Sequence of words → Embedding layer → LSTM

Image → Image feature vector 2,048 → Dense layer

LSTM, Dense layer → Dense layer → Softmax → Next word

Data Modelling : Train the Model(RNN-LSTM)

LSTM Model



05

Model Output & Deployment

Model output

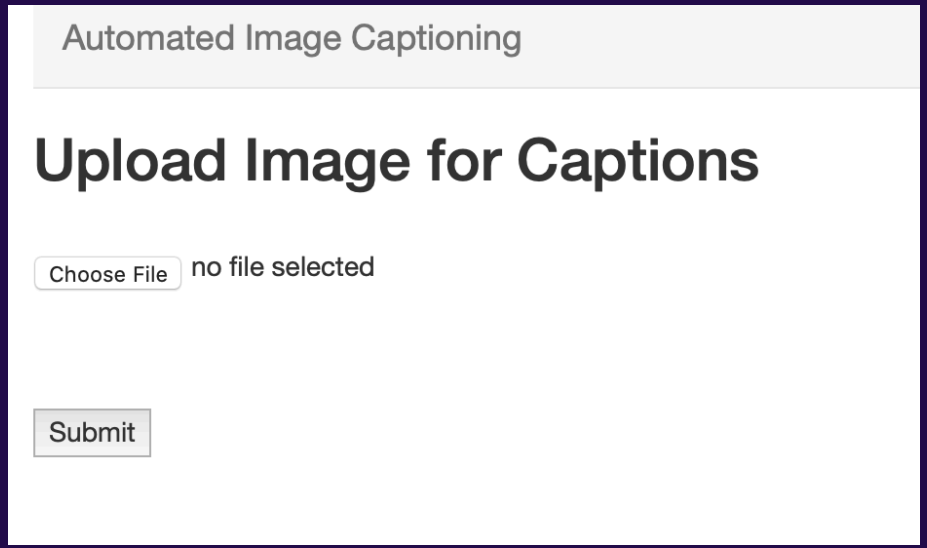
BLEU-1: 0.634139
BLEU-2: 0.404242
BLEU-3: 0.301511
BLEU-4: 0.167285

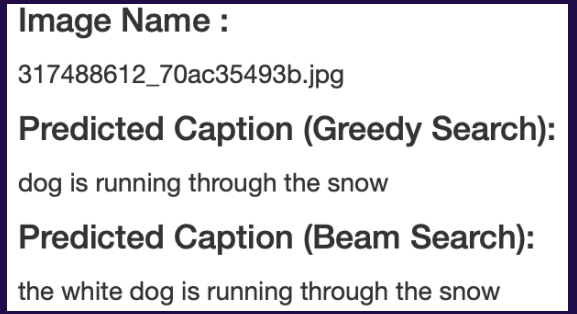
surfer is surfing in the water

dog is running on the beach

Deployment using Django

http://127.0.0.1:8000





06

LIMITATIONS/ RECOMMENDATIONS

LIMITATIONS

Model doesn’t perform well on cluttered images.

Predicted caption:
man in black shirt is standing in front of the camera

Output has repetitive words.

Predicted caption:
woman with red hair and white shirt and white shirt and white shirt and white shirt and white shirt

Generated captions are just wrong.

Predicted caption:
man is jumping over the top of the air

RECOMMENDATIONS

➢ *Train Model on large and diverse image datasets*	*MS COCO dataset (170K images)*
➢ *Try different pre-trained models for feature extraction*	*Xception, MobileNet*
➢ *Hyper-parameters tuning*	*Learning rate, Drop-out*
➢ *Use Attention Mechanism decoder*	*Relate words to a specified area of an image.*
➢ *Use powerful machines.*	

Questions??