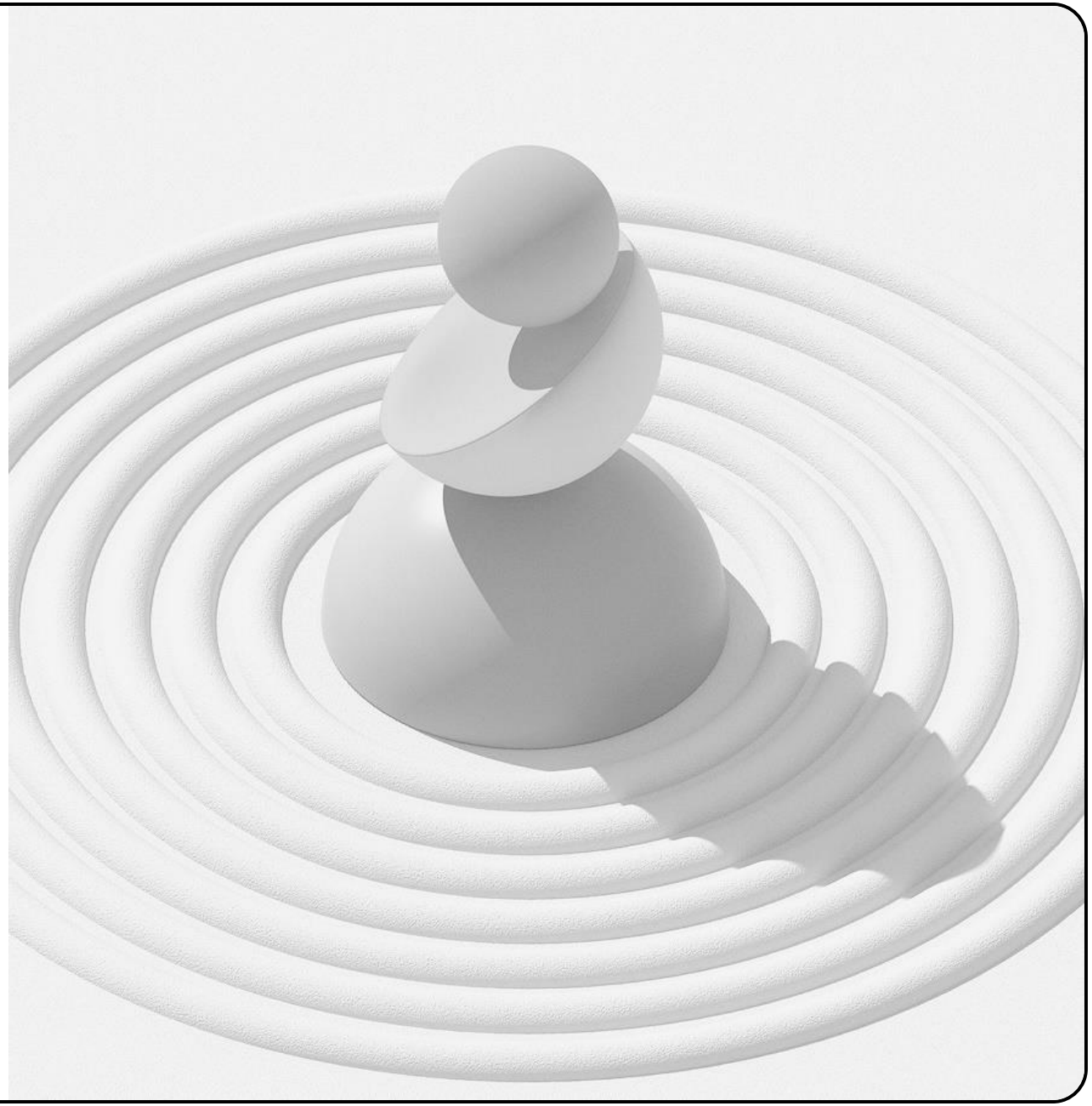
Presented by Pratik Davidson Deogam

PROJECT: IMAGE CAPTIONING

CONTENTS



Introduction

Image captioning is an AI task that takes an image as input and outputs a textual description of its content.

It combines two major fields:
1. **Computer Vision** → To understand objects, scenes, and actions in the image.
2. **Natural Language Processing (NLP)** → To generate grammatically correct and meaningful captions.

Dataset

Data - Mini COCO2014 Dataset for Image Captioning

Description:
Size – 3.06 GB
Images – 18k
JSON file – Contains image id and caption data
Total data – 93950
Training data: 75160
Validation data: 18790

Data Pre-processing

Image Data

1. Resize - (224x224x3)
2. Normalization - [0-255] → [0-1]

Caption Data

1. Vocabulary size - 3820
2. Tokenizer - WordPieceTokenizer

MODEL ARCHITECTURE

Image Encoder

- Vision Transformer (ViT)

Text Decoder

- Transformer Decoder

Parameters

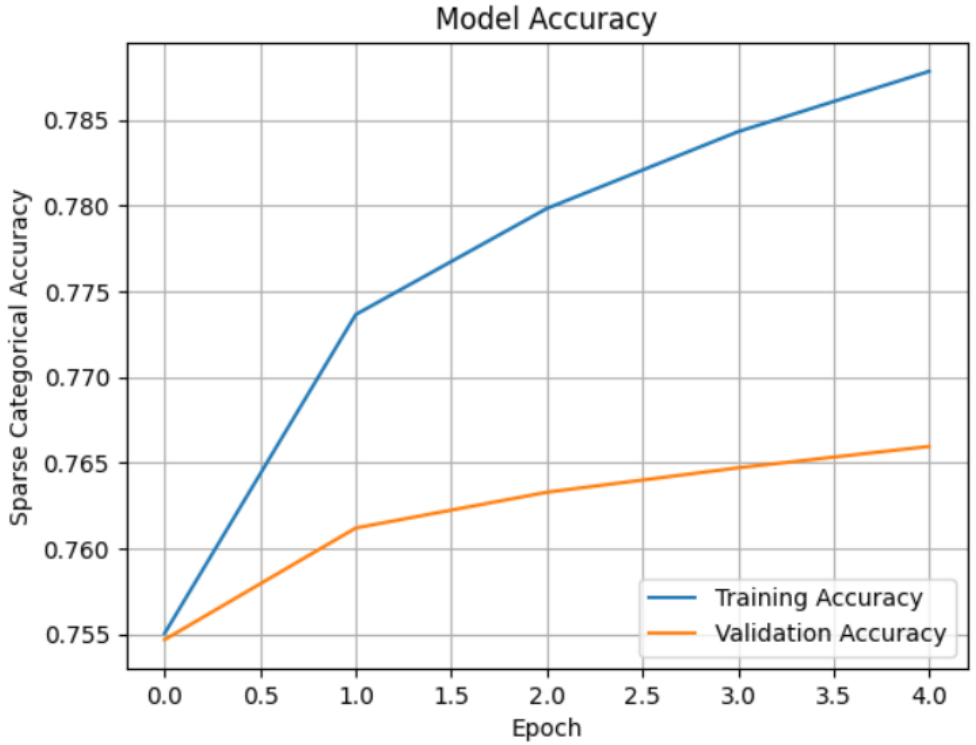
Total params: 90M

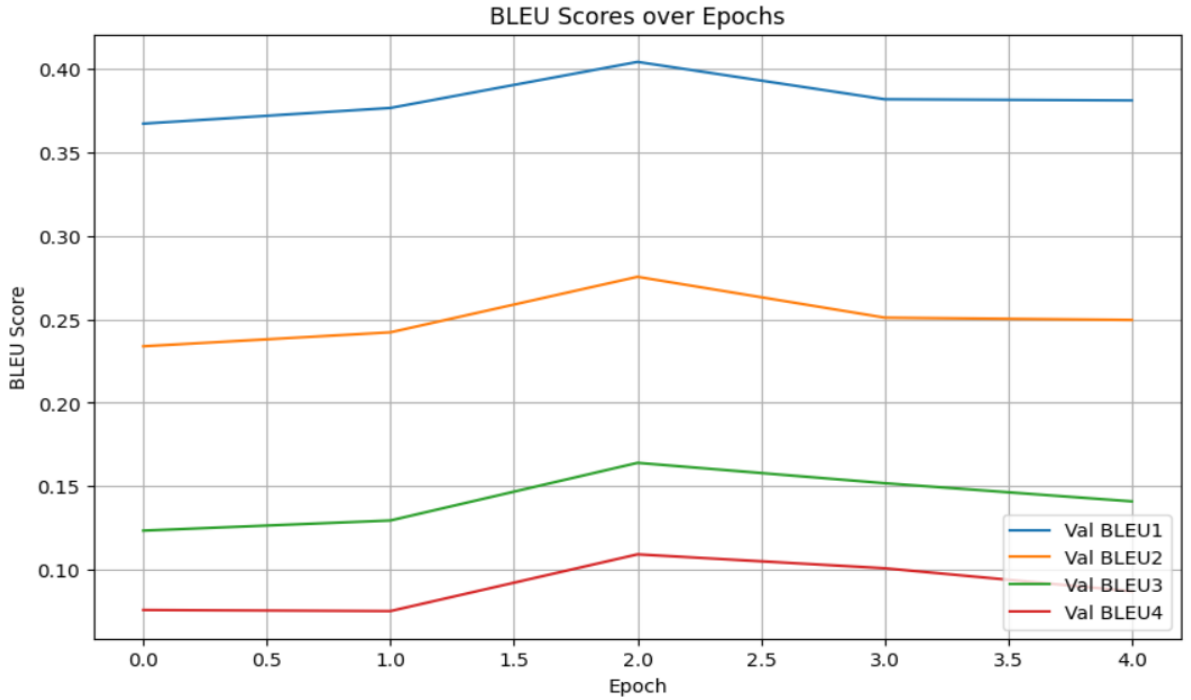
Trainable params: 4M

Non-trainable params: 86M

Layer (type)	Output Shape	Param #	Connected to
image (InputLayer)	(None, 224, 224, 3)	0	-
caption (InputLayer)	(None, None)	0	-
vi_t_backbone (ViTBackbone)	(None, 197, 768)	85,798,656	image[0][0]
token_and_position… (TokenAndPositionE…	(None, None, 256)	2,568,192	caption[0][0]
dense (Dense)	(None, 197, 256)	196,864	vi_t_backbone[0]…
transformer_decoder (TransformerDecode…	(None, None, 256)	790,784	token_and_positi… dense[0][0]
dropout_14 (Dropout)	(None, None, 256)	0	transformer_deco…
dense_1 (Dense)	(None, None, 3820)	981,740	dropout_14[0][0]

TRAIN/VAL RESULTS





Note: BLEU Score is calculated over 32 validation samples.

THANK YOU!