IMAGE CAPTIONING SYSTEM

Presented by: Roopsi Srivastava (202410116100173)

Rani Kumari(202410116100163)

Rabita Yadav(202410116100)

INTRODUCTION

•**What is Image Captioning?**

A process of generating textual descriptions for an image using both computer vision and natural language processing.

•**Why is it Important?**

•Helps visually impaired people

•Used in content-based image retrieval

•Powers features in social media and e-commerce platforms

Real-World Applications

- **Examples:**
- Google Photos automatic captions
- Facebook alt text for visually impaired users
- Pinterest visual search
- Autonomous vehicles understanding their surroundings
- Add 2–3 image examples with captions for effect

How Image Captioning Works

- **High-Level Overview:**
  - **Image Input** → CNN (extract features)
  - **Feature Vector** → RNN/LSTM (generate sentence)
  - **Output** → Caption
- **Visual Aid:** Show a diagram of the above flow.

Components Involved

- **1. CNN (Convolutional Neural Networks):**
  Extracts image features (e.g., using VGG, ResNet)
- **2. RNN/LSTM/GRU (Language Model):**
  Generates sequence of words based on extracted features
- **3. Attention Mechanism (Optional):**
  Focuses on specific parts of the image while generating each word

Algorithms & Models

- **Encoder-Decoder Architecture**
- **Encoder:** CNN
- **Decoder:** RNN or LSTM with Word Embeddings
- **Pretrained Models:**
- InceptionV3, ResNet50 (for image encoding)
- Beam Search or Greedy Decoding (for caption generation)

Tools and Technology

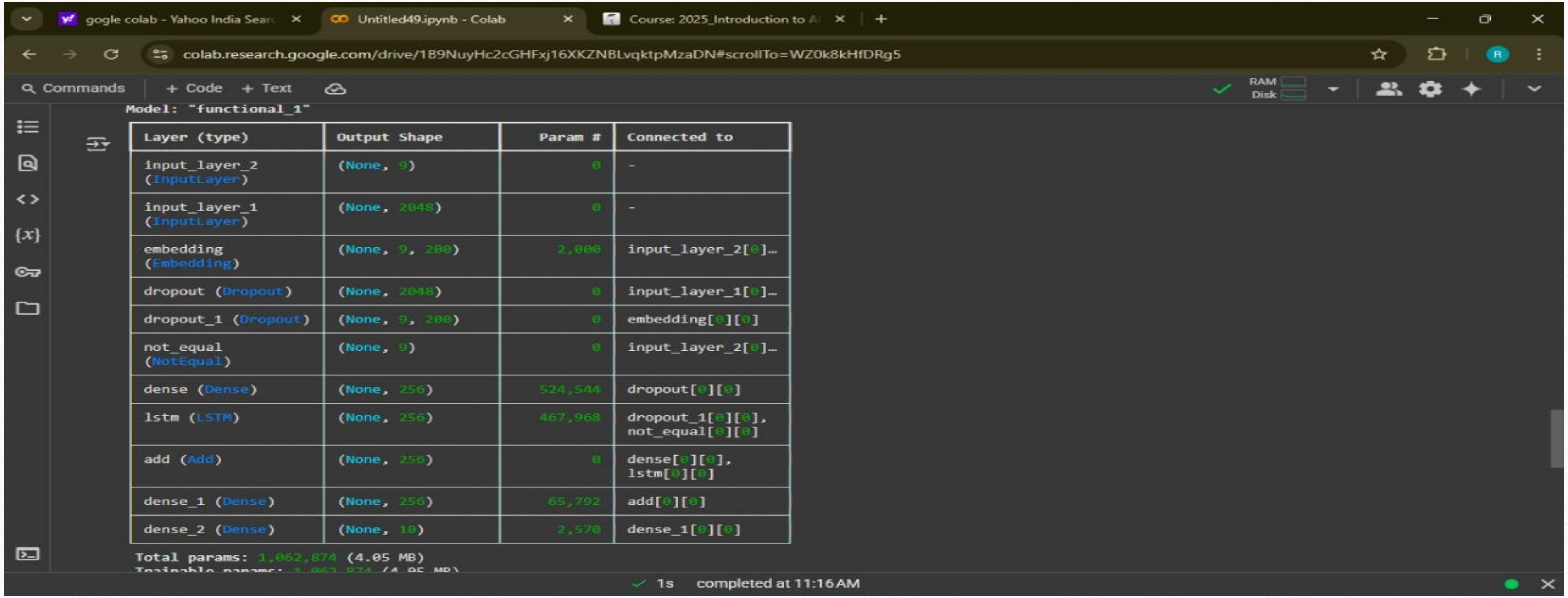
- **Languages:** Python
- **Libraries:** TensorFlow / PyTorch, OpenCV, NLTK
- **Frameworks:** Keras, HuggingFace Transformers
- **Other:** Google Colab

Future Scope

- Improved caption quality with GPT models
- Multilingual caption generation
- Real-time captioning in AR/VR
- Personalized image descriptions

Image Uploaded

Output


gogle colab - Yahoo India Searc
Untitled49.ipynb - Colab
Course: 2025_Introduction to A
colab.research.google.com/drive/1B9NuyHc2cGHFxj16XKZNBLvqktpMzaDN#scrollTo=WZ0k8kHfDRg5
Commands
+ Code
+ Text
RAM
Disk
R
1s
completed at 11:16 AM

```
Model: "functional_1"
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━┓
┃ Layer (type)        ┃ Output Shape      ┃    Param # ┃ Connected to        ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━┩
│ input_layer_2       │ (None, 9)         │          0 │ -                   │
│ (InputLayer)        │                   │            │                     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ input_layer_1       │ (None, 2048)      │          0 │ -                   │
│ (InputLayer)        │                   │            │                     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ embedding           │ (None, 9, 200)    │      2,000 │ input_layer_2[0]…   │
│ (Embedding)         │                   │            │                     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ dropout (Dropout)   │ (None, 2048)      │          0 │ input_layer_1[0]…   │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ dropout_1 (Dropout) │ (None, 9, 200)    │          0 │ embedding[0][0]     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ not_equal           │ (None, 9)         │          0 │ input_layer_2[0]…   │
│ (NotEqual)          │                   │            │                     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ dense (Dense)       │ (None, 256)       │    524,544 │ dropout[0][0]       │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ lstm (LSTM)         │ (None, 256)       │    467,968 │ dropout_1[0][0],    │
│                     │                   │            │ not_equal[0][0]     │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ add (Add)           │ (None, 256)       │          0 │ dense[0][0],        │
│                     │                   │            │ lstm[0][0]          │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ dense_1 (Dense)     │ (None, 256)       │     65,792 │ add[0][0]           │
├─────────────────────┼───────────────────┼────────────┼─────────────────────┤
│ dense_2 (Dense)     │ (None, 10)        │      2,570 │ dense_1[0][0]       │
└─────────────────────┴───────────────────┴────────────┴─────────────────────┘
 Total params: 1,062,874 (4.05 MB)
 Trainable params: 1,062,874 (4.05 MB)
```

Thank
You