CS 613: Natural Language Processing

IIT Gandhinagar

Sem-I - 2025-26

ASSIGNMENT - 1

Team Name- Lingo Limbos (Team 6)
Team Members- Abhay, Abhiroop, Harsh, Maharshi, Mithlesh, Saravanan, Uday S.G.
Our Experiment Github Link: GitHub

I. Report

1. Introduction

Motivation :

Image captioning bridges the gap between visual and linguistic modalities by generating an image description in natural language that reflects the contents of a given input image. It has made significant progress in recent years, yet current models focus on image description based on what is present in the image, not considering contextual information. This makes the approach inefficient for tasks requiring deep understanding in real-world scenarios, such as education and healthcare. Traditional models often hallucinate details and lack contextual relevance, providing misinformation that has real-world consequences. By adding external knowledge sources, the accuracy and relevance of the description can be improved a lot. In the future, such systems could support conversational agents, virtual guides, or automated content creation tools that speak multiple languages and maintain factual integrity.

Relation with NLP :

The problem involves retrieving, understanding and generating contextually relevant textual descriptions based on both structured and unstructured sources. The model needs to process the retrieved text, aligning it with visual input and producing factually correct descriptions in multiple languages. This involves tasks such as language modelling, information retrieval, named entity recognition and fact validation and verification, which are all fundamental tasks of natural language processing.

Problem type :

This is primarily a problem of text generation, also involving elements of information retrieval. In addition to that, because we are dealing with Indic languages, the other aspects it includes are knowledge integration and multilingual translation.

2. Related Work

State-of-the-art

For this problem, the current best approach is the DIR model, proposed by Hao Wu and colleagues in the paper "DIR: Retrieval-Augmented Image Captioning with Comprehensive Understanding". DIR focuses on improving image captioning by introducing diffusion-guided image representation to make the image encoder capture a broader range of semantic details, such as objects, actions, and scene context. It builds a comprehensive retrieval database that decomposes textual knowledge into fine-grained attributes and retrieves relevant details to support the captioning process. This design helps the model generalise better to novel or out-of-domain scenarios while maintaining strong in-domain performance. Importantly, these improvements come without adding inference overhead, making DIR a robust and efficient solution. However, due to the unavailability of code for this DIR model and considering that DIR and other state-of-the-art models are built on SmallCap, we are implementing SmallCap model.

Baseline implementations availability

No, the baseline implementations are not available. We have codes for SmallCap, RobustCap. But we were unable to get RobustCap working. So, we used the model for the SmallCap available on SmallCap.

Table of results from a previous paper

The paper shows that DIR is able to perform better than existing models, especially when it comes to captioning images with unseen objects. Attached the table below:

Method	Train		Flickr30k		NoCaps Val			
	Data	Para.	Test		Out-domain		Overall	
			C	S	C	S	C	S
Heavyweight training methods								
BLIP [17]	129M	446B	-	-	115.3	14.4	113.2	14.8
BLIP2 [18]	129M	1.2B	-	-	124.8	15.1	121.6	15.8
ViECap [12]	COCO	124M	47.9	13.6	65.0	8.6	66.2	9.5
Lightweight training methods								
MiniGPT4 [40]	5M	3.94M	78.4	16.9	110.8	14.9	108.8	15.1
ClipCap [24]	COCO	43M	-	-	49.1	9.6	65.8	10.9
SmallCap [29]	COCO	7M	60.6	-	-	-	-	-
EVCap [19]	COCO	3.97M	84.4	18.0	116.5	14.7	119.3	15.3
DIR (ours)	COCO	3.97M	**85.7**	**18.2**	**118.7**	**15.1**	**120.7**	**15.5**

Table 1. Out-of-domain performance comparison on Flickr30k and NoCaps (out-domain and overall). The best performance among lightweight methods is highlighted in bold, with a light blue background added when it also surpasses all heavyweight models.

3. Datasets

a. Is the dataset available?

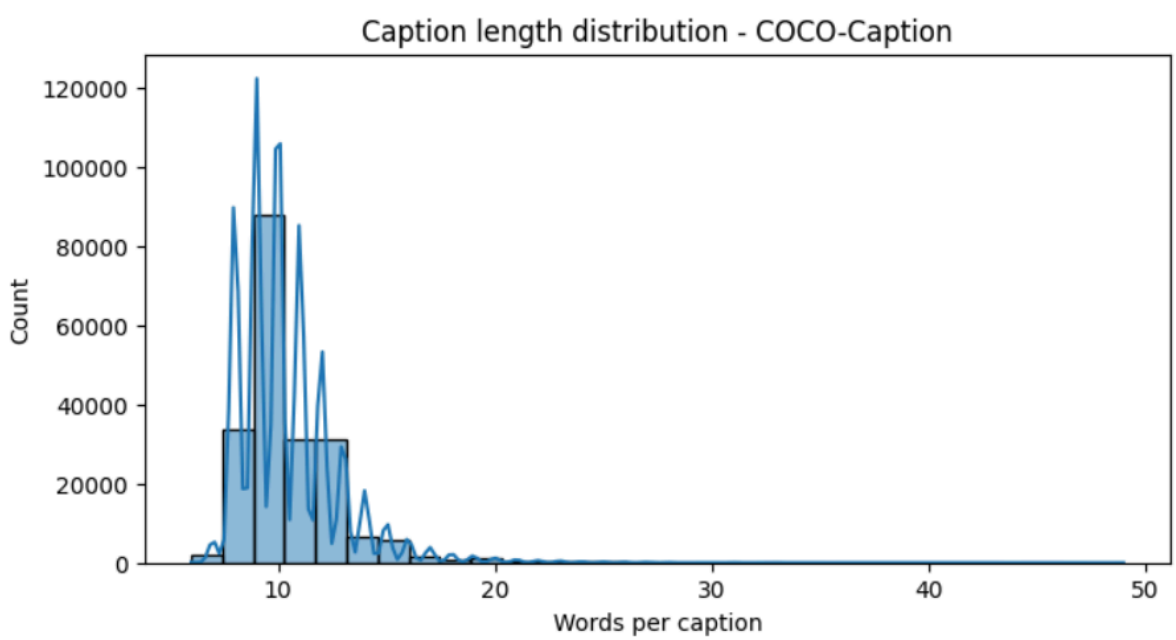
Yes, DIR uses the COCO dataset mainly - COCO, which is available open source.

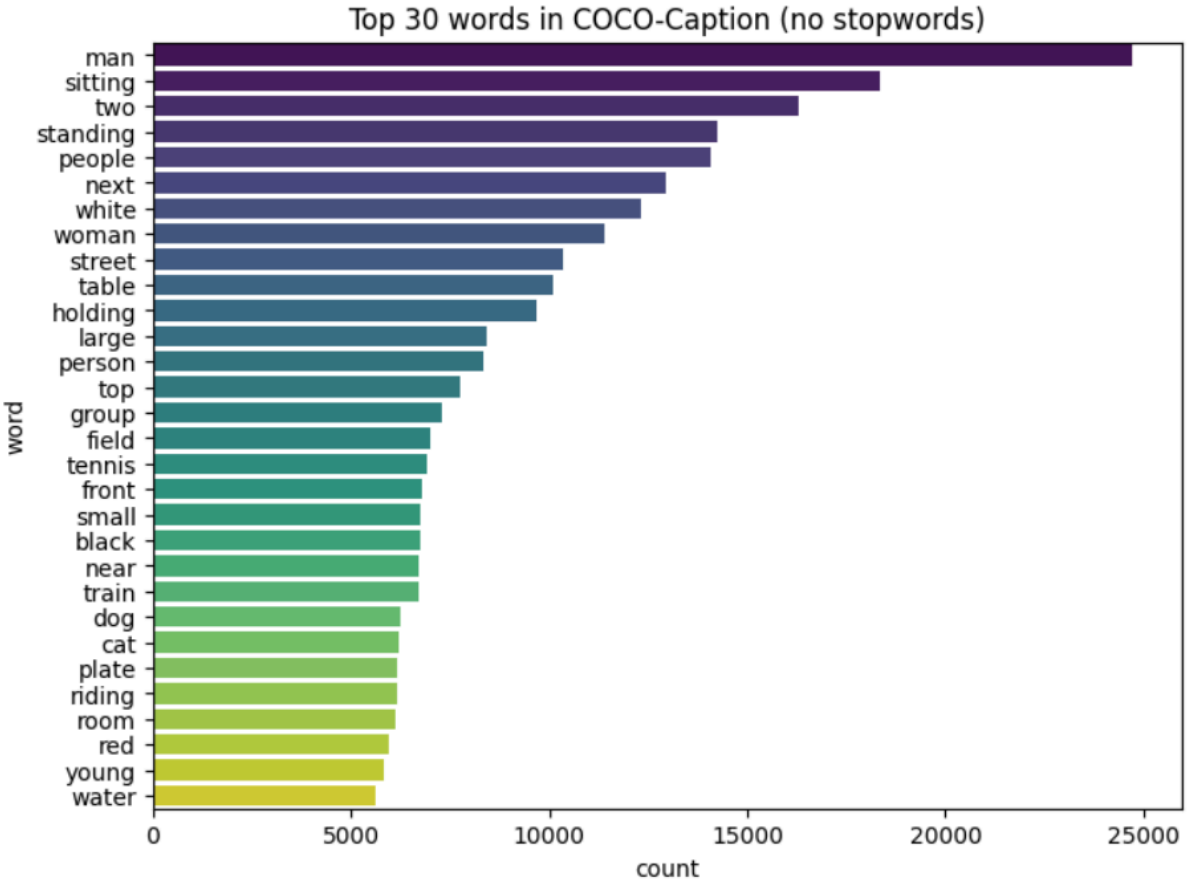
COCO:

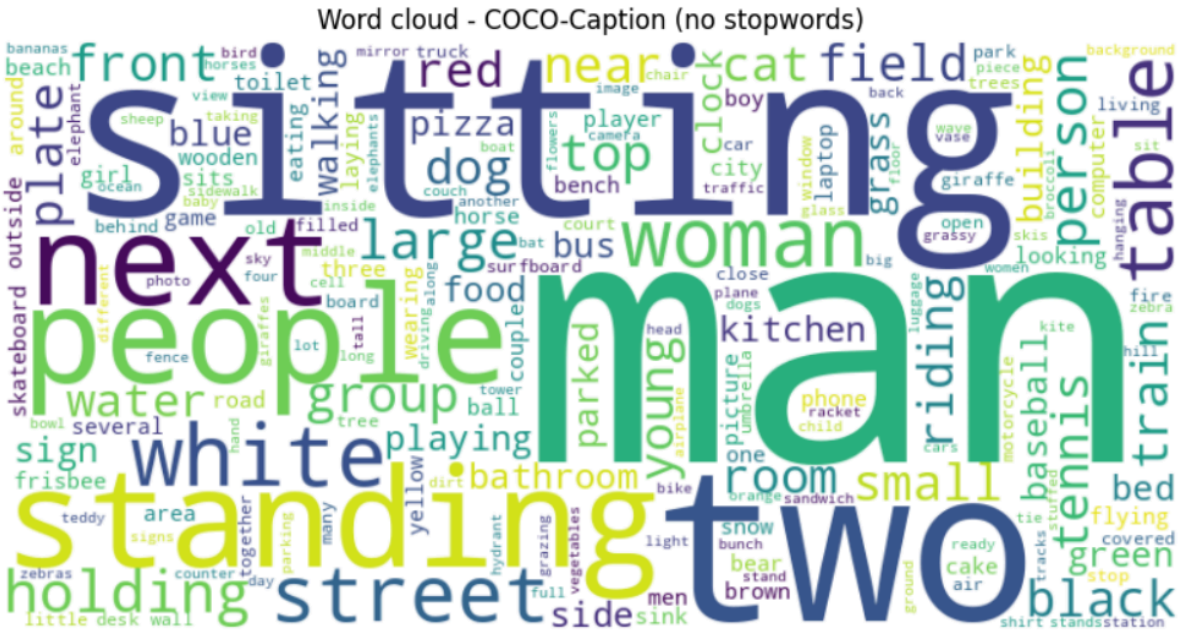
```
Dataset size: 243429 captions
Images: 40505
Captions: 243429
Average captions per image: 5.00
```

Vocabulary size: 17039

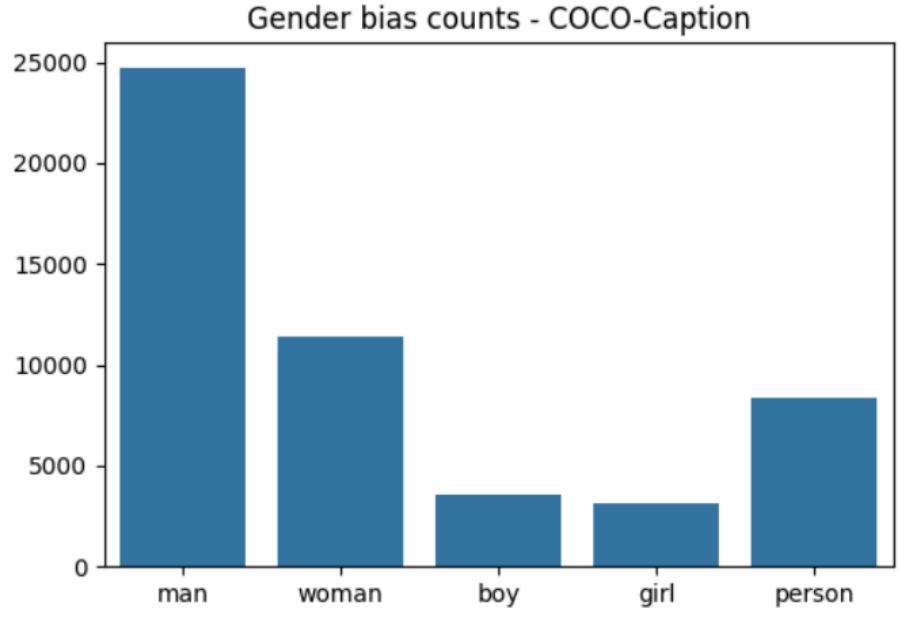
Vocab_size and len(words) ratio: 0.0079


Caption length distribution - COCO-Caption
Count
120000
100000
80000
60000
40000
20000
0
10
20
30
40
50
Words per caption

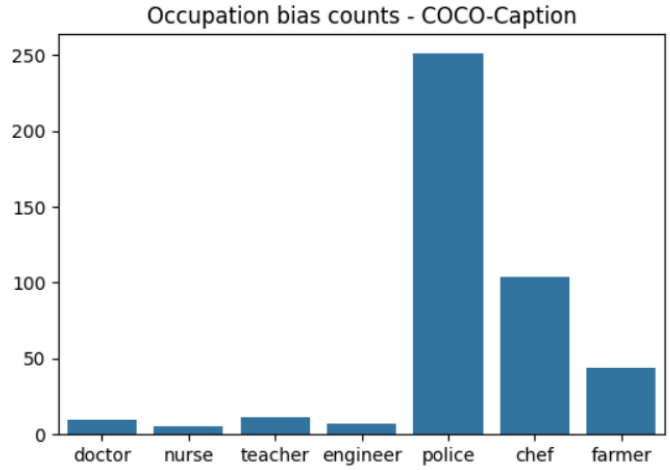

Top 30 words in COCO-Caption (no stopwords)
word
man
sitting
two
standing
people
next
white
woman
street
table
holding
large
person
top
group
field
tennis
front
small
black
near
train
dog
cat
plate
riding
room
red
young
water
0
5000
10000
15000
20000
25000
count



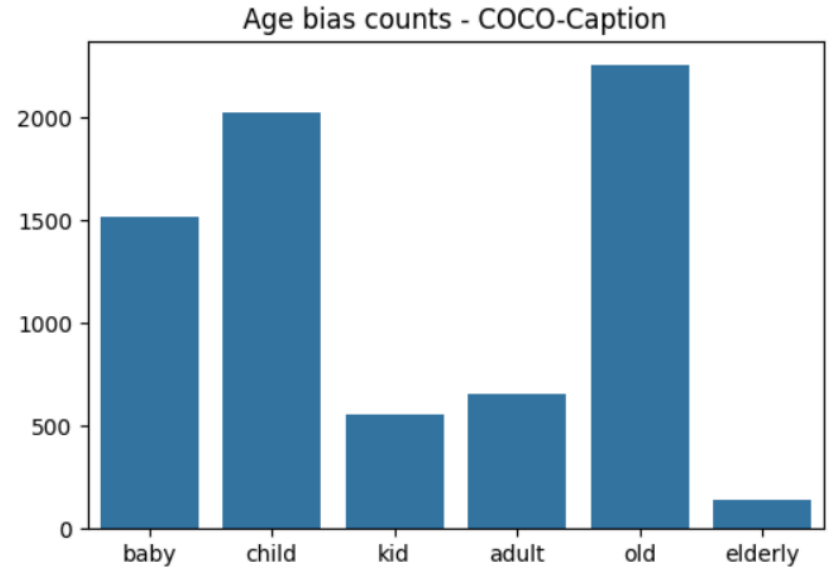
Gender mentions: {'man': 24718, 'woman': 11418, 'boy': 3547, 'girl': 3176, 'person': 8331}



```
Occupation mentions: {'doctor': 10, 'nurse': 5, 'teacher': 11, 'engineer': 7, 'police': 251, 'chef': 104, 'farmer': 44}
```




```
Age mentions: {'baby': 1518, 'child': 2028, 'kid': 553, 'adult': 653, 'old': 2258, 'elderly': 140}
```



4. Experimental Plan (10%)

a. What experimental setting will you choose?

- For the baseline, we will reproduce results from the reference paper (e.g., SmallCap or other SOTA baselines). And we are planning to branch in 2 directions- Architecture modification, Multilingual extension
- For the architecture modification, we will experiment with changes to encoder-decoder pipelines, attention mechanisms, or vision-language fusion modules.

- For the multilingual, we will extend the baseline model with multilingual embeddings.

i. How to do a train/development/test split?

- It is already split as the COCO dataset is what we are using.

ii. How to find the best params for your algos and baselines?

- We use random search followed by Bayesian optimisation (e.g., Optuna), since this is far more efficient for large models like image captioning and helps find the best parameters with fewer experiments.

b. What metric will you choose? And why? Does this relate to the problem you are solving?

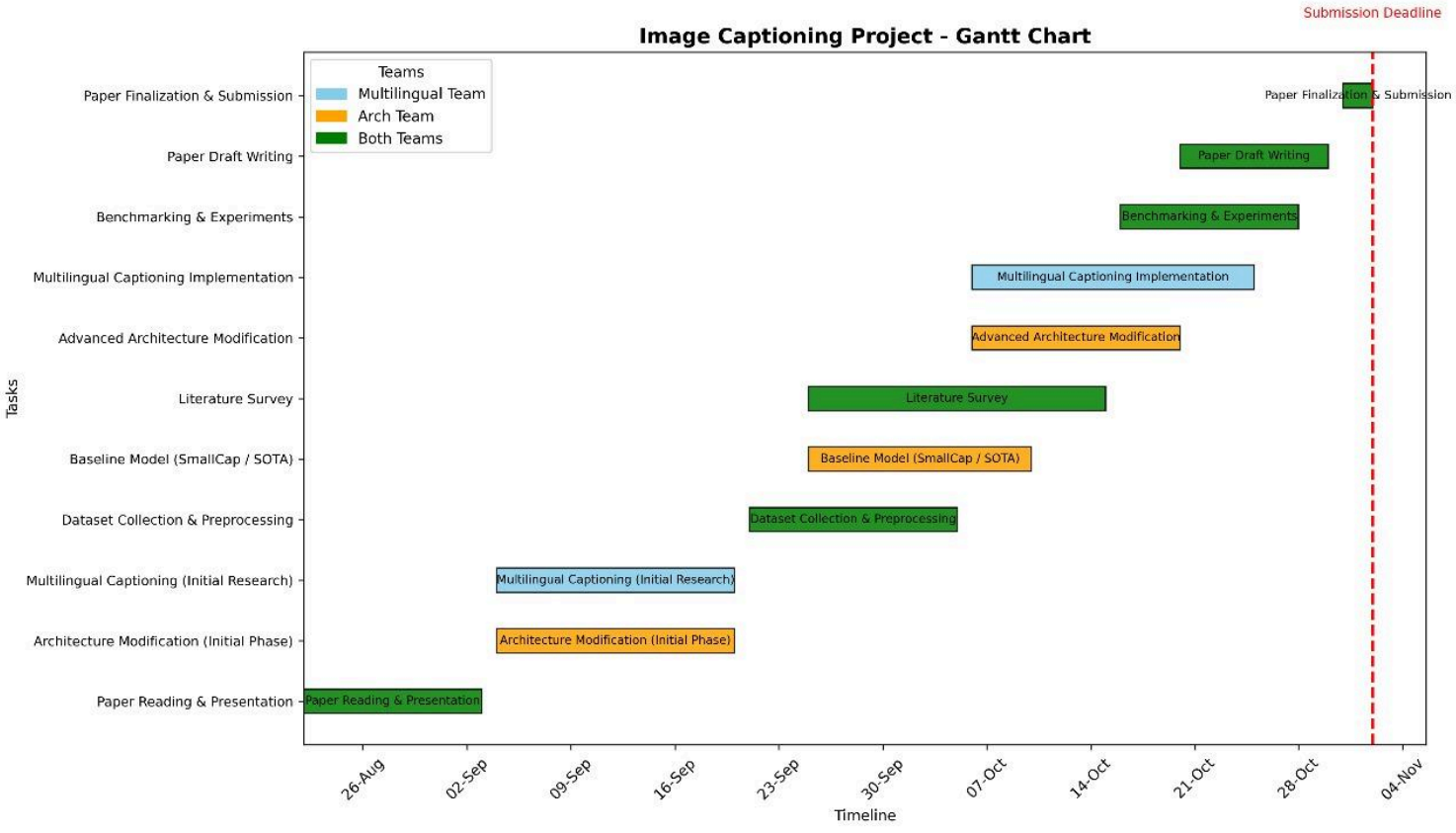
- BLEU score, METEOR, ROUGE_L, CIDEr, SPICE. Each metric mentioned above tells about a different semantic meaning in the output. And importantly, the current SOTA papers use these metrics, so we use these for comparison.

c. What level of system effort will you require? Please paint a scenario for the final demo that you will present. Will it be a live website? Or, an Android app?

- A moderate system effort is required: a single GPU (e.g., RTX 3090 or V100) with 16–32 GB RAM and ~500 GB storage is sufficient to deploy modified architecture models, while the multilingual extension can be trained sequentially on the same GPU. We will be showing a live website made using Streamlit.

5. Project Management (10%)

a. Please create a Gantt Chart. Here is a link on how to do so in Python [0, 1, 2].



b. **Computation (RAM, GPUs, CPU Cores, Hard Disk, etc.) resources are needed.**
   - a single GPU (e.g., RTX 3090 or V100) with 16–32 GB RAM and ~500 GB storage

c. **When will you consider the project a success?**
   - The project will be considered a success when we are able to publish our work.

d. **What is the biggest risk that will lead to project failure, and how will you address it?**
   - The biggest risk is that our architecture changes might not give better results than the baseline. To handle this, we've split the team into two: one group focuses on tuning the architecture, while the other works on multilingual captioning. This way, even if the architecture improvements don't work as expected, we still have the multilingual part as a novel contribution for the project and potential publication.

e. **How will you split up the project amongst yourselves? What tasks are distributed amongst you?**
   Group-1 Pre-Processing on Multilingual Dataset: Harsh & Maharshi & Uday
   Group-2 Mithlesh & Saravanan(Pipeline Finetuning), Abhay & Abhiroop(Architecture Modifications)

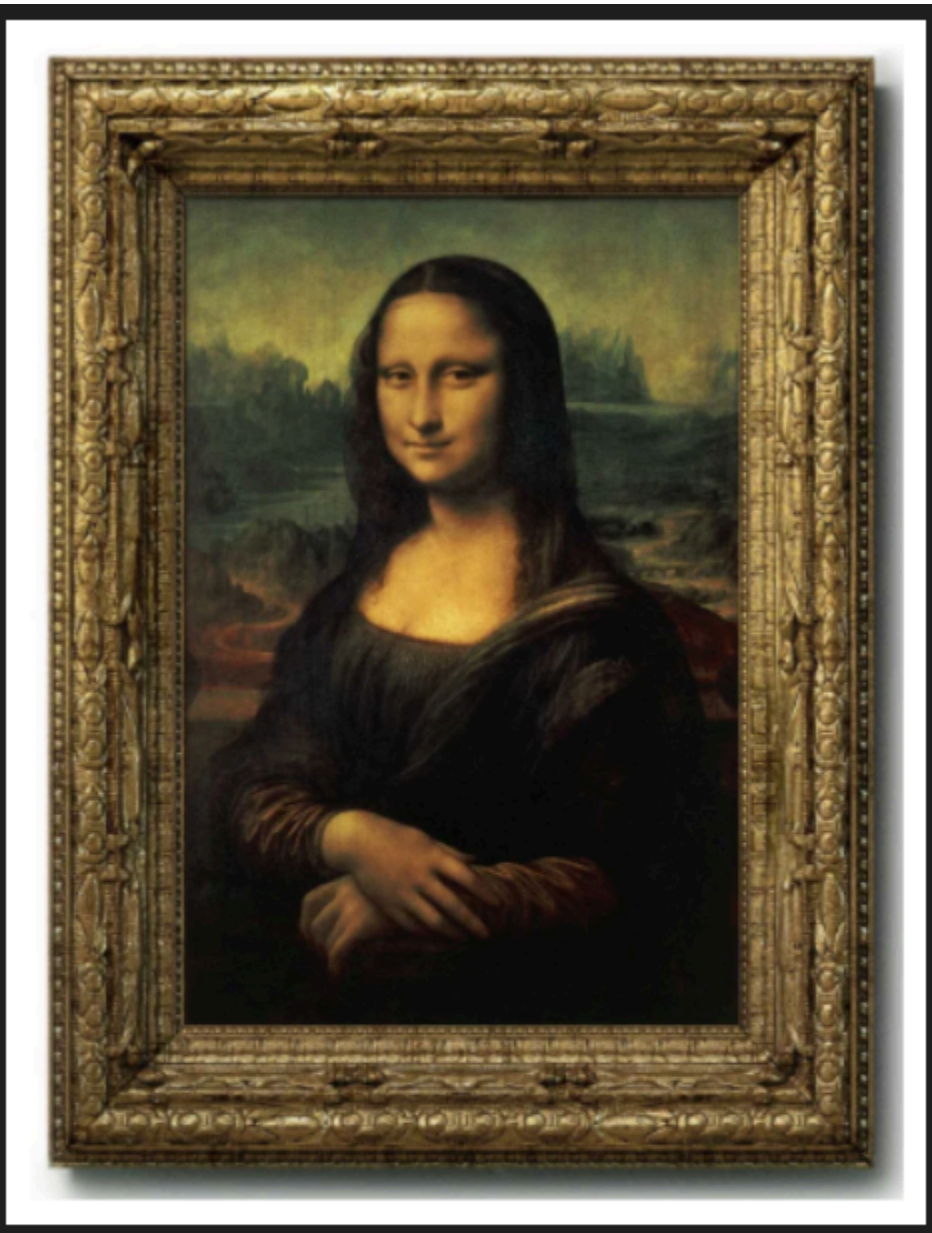
II. Experiment

Our Experiment Github Link: GitHub

Example Results on SmallCap

```
Using device: cpu
Loading GPT2 tokenizer...
Loading SmallCap model...
✅ SmallCap loaded
Loading template...
✅ Template loaded
Loading CLIP feature extractor...
✅ Image loaded: C:\Users\sarav\Downloads\monalisa.jpg
✅ Prompt prepared
✅ Caption generated in 4.61 seconds
```

Example 1:



```
Prompt:
 This image shows

Generated caption:
 a picture of a painting
```

Example 2:

```
Prompt:
 This image shows

Generated caption:
 a large building with a tree in the background
```

Example 3:

```
Prompt:
 This image shows

Generated caption:
 a view of a train track
```

Results on COCO Dataset

```
loading annotations into memory...
Done (t=0.00s)
creating index...
index created!
Loading and preparing results...
DONE (t=0.00s)
creating index...
index created!
tokenization...
setting up scorers...
computing Bleu score...
{'testlen': 771, 'reflen': 886, 'guess': [771, 671, 571, 471], 'correct': [458, 215, 87, 34]}
ratio: 0.8702031602698982
Bleu_1: 0.512
Bleu_2: 0.376
Bleu_3: 0.265
Bleu_4: 0.184
computing METEOR score...
METEOR: 0.172
computing Rouge score...
ROUGE_L: 0.407
computing CIDEr score...
CIDEr: 0.556
computing SPICE score...
SPICE: 0.118
Bleu_1: 0.5117
Bleu_2: 0.3758
Bleu_3: 0.2647
Bleu_4: 0.1843
METEOR: 0.1718
ROUGE_L: 0.4072
CIDEr: 0.5559
SPICE: 0.1185
```

- As we can observe, the Bleu score is maximum at BLEU-1 (about 51% of unigrams match) and decreases with the longer n-gram. This indicates that the captions are capturing words but not longer sequences precisely.
- Rouge_L score measures the longest common subsequence. A nearly 41% overlap shows moderate structural similarity with the reference captions.
- SPICE score indicates the scene graph semantics. A low spice score indicates that relationships are learned poorly.
- CIDEr uses TF-IDF weighting for words. A nearly 0,56 score indicates the model captures important content moderately well.

References:

1. SMALLCAP: Lightweight Image Captioning Prompted with Retrieval Augmentation
2. DIR: Retrieval-Augmented Image Captioning with Comprehensive Understanding
3. Understanding Retrieval Robustness for Retrieval-Augmented Image Captioning
4. COCO-Caption

Acknowledgement:
Grammarly for checking Grammar