Enhancement of Compositional Prompts and their fine-tuning using Blip-2

Semester Project of Generative AI & LLM Course (Fall, 2024)

Group Members:

Hassan Javaid (MSCS23001)
Sauda Maryam (MSDS22025)
Javeria Saeed (MSCS23010)

Contents

Project Introduction

- **An Overview**: Compositional learning uses **state/attribute**, and **object** to define or describe an image e.g. "a picture of an **acrylic shoe**". This compositional prompt uses one attribute to describe its object. Can we enhance this prompt for inclusion of multi-attributes?

- **Problem Statement:** To enhance and redefine the compositional prompt to include multiple attributes for the given object.

- **Prompt Enhancement and redefinition:**
  - <u>Original Prompt:</u> "a picture of an **acrylic shoe**"
  - <u>Enhanced Prompt:</u> "a picture of an **acrylic shoe** with **white** color and **rubber** material".

A Brief Overview: Compositional Learning

- **Definition**: Learning complex systems by composing them from simpler, reusable components or concepts. Enables systematic generalization and hierarchical reasoning. In short, we can define object with single or multiple attributes/states.
- **Key Features:**
  - **Modularity**: Reusable components (e.g., "**red**" + “**marker**" = "**red marker**").
  - **Systematic Generalization**: Applies concepts to unseen combinations.
  - **Hierarchical Structure**: Builds abstractions from simpler elements.
- **Applications:** Generative AI, NLP.

A Brief Overview: Surveyed Papers / Strategies

- **Paper-1:** Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Li, J., Li, D., Savarese, S. and Hoi, S. (ICML, 2023)

- **Paper-2:** Troika: Multi-path cross-modal traction for compositional zero-shot learning. Huang, S., Gong, B., Feng, Y., Zhang, M., Lv, Y. and Wang. (CVPR, 2024)

- **Paper-3:** Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. Lu, X., Guo, S., Liu, Z. and Guo, J. (CVPR, 2023)

- **Paper-4:** Learning transferable visual models from natural language supervision. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G. (ICML, 2021)

Dataset Overview - UT-Zappos50K

› **Description: UT-Zappos50K** is a large dataset for fine-grained visual categorization and attribute-based learning, focusing on footwear images.

› **Key Features:**

– **Size:** ~50,000 images of shoes.

– **Categories:** Includes multiple types like sandals, sneakers, boots, etc.

– **Attributes:** Each image is annotated with fine-grained attributes:

- **Colors:** Red, blue, black, white, etc.
- **Materials:** Leather, canvas, rubber, synthetic, etc.
- **Closure:** Lace-up, slip-on, buckle, velcro.
- **Heel Height:** Flat, low, medium, high.

UT-Zappos50K - Dataset Overview

› **Applications:**

- Compositional learning (e.g., combining attributes like "red leather boots").
- Zero-shot learning for unseen combinations of attributes.

› **Reference website: [UT-Zappos50K Dataset]**

› **Sample Images:**



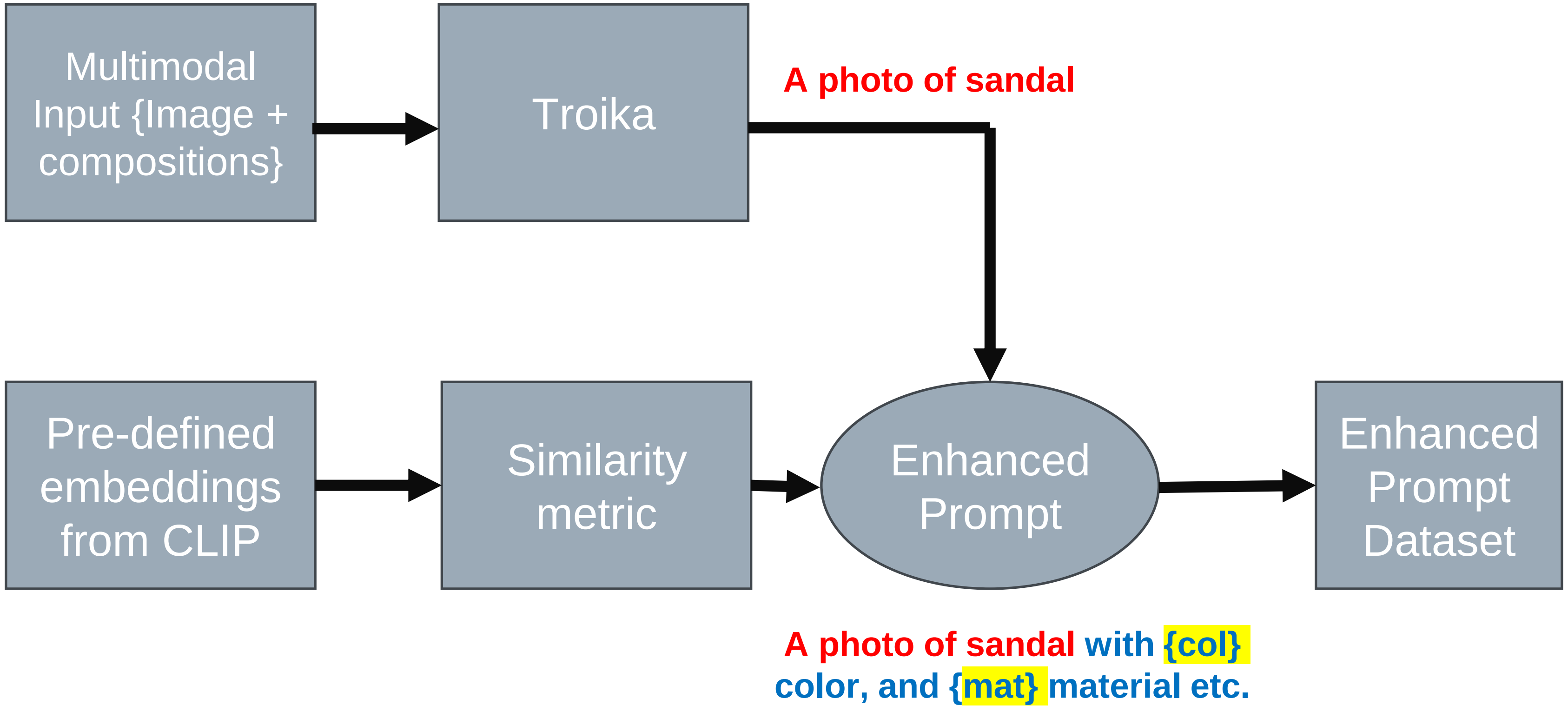

Training Methodology

- **Training Steps Outline:**
  - **1st Step:** Creation of enhanced prompt datasets using similarity metric calculation with pre-defined CLIP embeddings
    - **Color types** = {black, white, brown, red}
    - **Material** = {leather, fabric, rubber, synthetic}
  - **2nd Step:** Using Parameter Efficient Fine-tuning (PEFT) to implement fine-tuning on Blip-2 pre-trained model to generate enhanced prompts
    - Used **Low-Rank Adaptation (LoRA)** for implementing PEFT

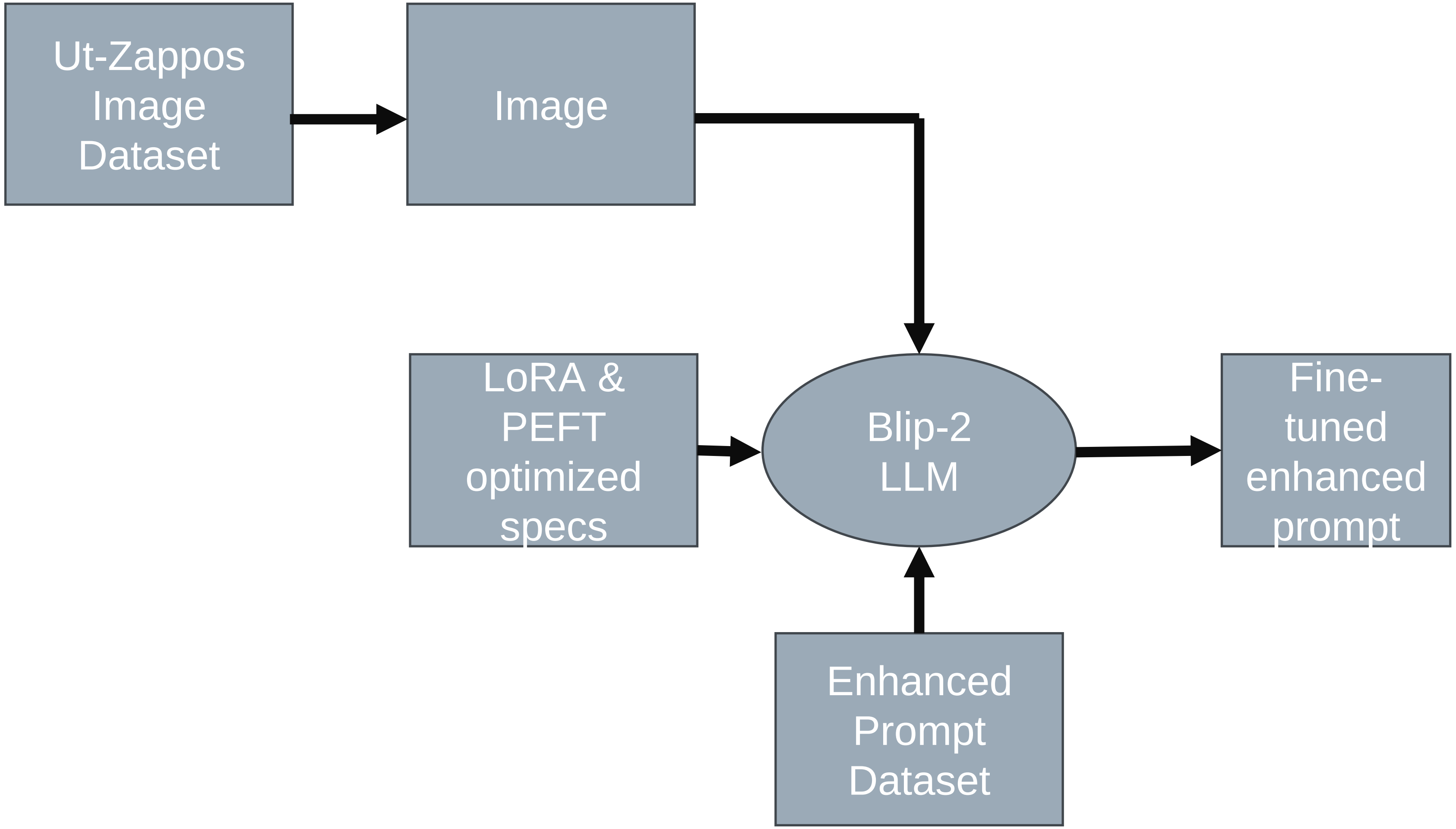
Training Pipeline & Specifications

- **Original prompt** generated from *Troika*
- **CLIP embeddings** calculated on pre-defined color and material types
  - **Color types** = {black, white, brown, red}
  - **Material** = {leather, fabric, rubber, synthetic}
- **Creation of enhanced prompts dataset** for around 24k Ut-Zappos shoe images (Fine Tuning done on 1k images)
  - A picture of {**Canvas_Shoes.Boat.Shoes**} with {**col**} color and {**mat**} material
- **Using LORA** for implementing Parameter Efficient Fine-Tuning (PEFT)
  - **Rank (r) = 16**
  - Scaling factor = 32
  - **Dropout = 0.1**
  - Target-modules = [q_proj, v_proj]
- **Training & PEFT of BLIP-2** using enhanced prompts

Training Architecture – 1st step



Training Architecture – 2nd step



A photo of sandal with {col} color, and {mat} material etc.

Training Results

› Untrained LLM:

```
# Generate caption
caption = generate_caption(t_img_path, untrained_model, processor, device)
print(f"Generated Caption: {caption}")
```

```
Expanding inputs for image tokens in BLIP-2 should be done in processing. Please follow ins
Generated Caption: a brown and tan boat shoe with a white sole
```

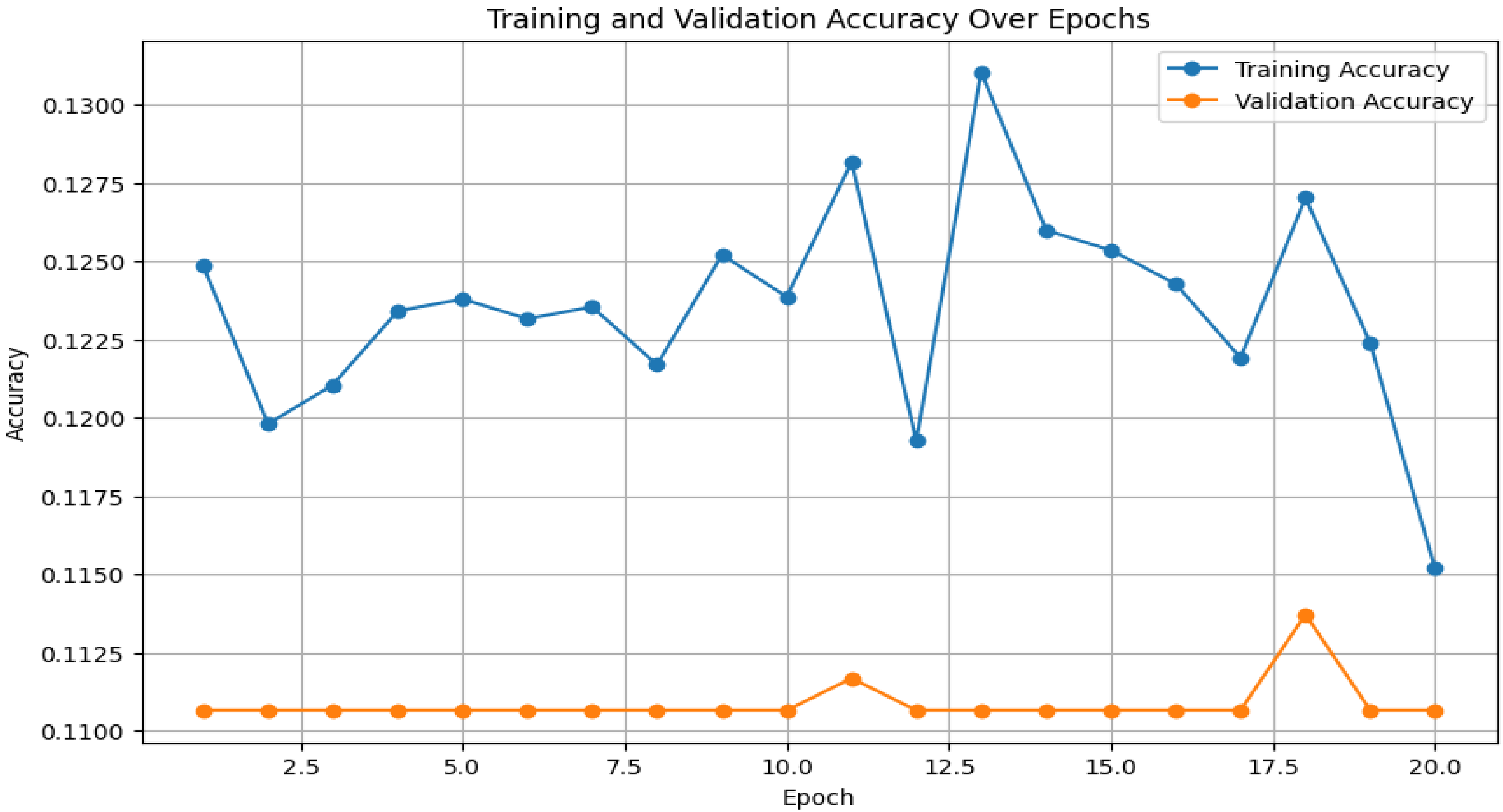
› Trained LLM:

```
# Generate caption
caption = generate_caption(t_img_path, model, processor, device)
print(f"Generated Caption: {caption}")
```

```
Generated Caption: A photo of Canvas_Shoes.Boat.Shoes with brown color and rubber material
```

Training Results

Training Results



Evaluation Scheme

- **Brief Outline:**
  - **For inference following inputs were given to the trained LLM**
    - Ground Truth caption
    - Image
  - **Output:** Generated caption
  - **Train-valid-test split =** 80-10-10
  - **Metrics:** BLEU, ROGUE, Semantic Similarity, CLIP (Semantic Similarity)
  - **Libraries used:** NLTK

Evaluation Metrics

- To measure the efficacy of our finetuning process, we used the following metrics for evaluation of generated prompts with ground-truth prompts.
  - **BLEU Score (Bilingual Evaluation Understudy):** Measures how closely machine-generated text matches reference text by comparing overlapping n-grams adjusted by a brevity penalty.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

    - Where **BP = Brevity Penalty**, accounts for shorter generated sentences

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$

    - c: length of candidate text, r: length of reference text.
    - p_n: Precision of n-grams.
    - w_n: Weight assigned to each n-gram (usually $w_n = \frac{1}{N}$)

Evaluation Metrics

– **ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation):** Measures recall and precision of n-grams or longest common subsequence.

› For **ROUGE-N** (n-gram based recall):

$$ROUGE - N = \frac{|\text{Overlapping n-grams}|}{|\text{Total n-grams in reference}|}$$

› For **ROUGE-L** (Longest Common Subsequence):

$$ROUGE - L = \frac{LCS(\text{candidate, reference})}{\text{length of reference}}$$

› Where LCS is the Longest Common Subsequence.

Evaluation Metrics

– **Similarity Score:** Quantifies the semantic similarity between two texts using cosine similarity of their vector representations.

$$\text{Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| ||\mathbf{B}||}$$

Evaluation Results – Best and Worst Cases

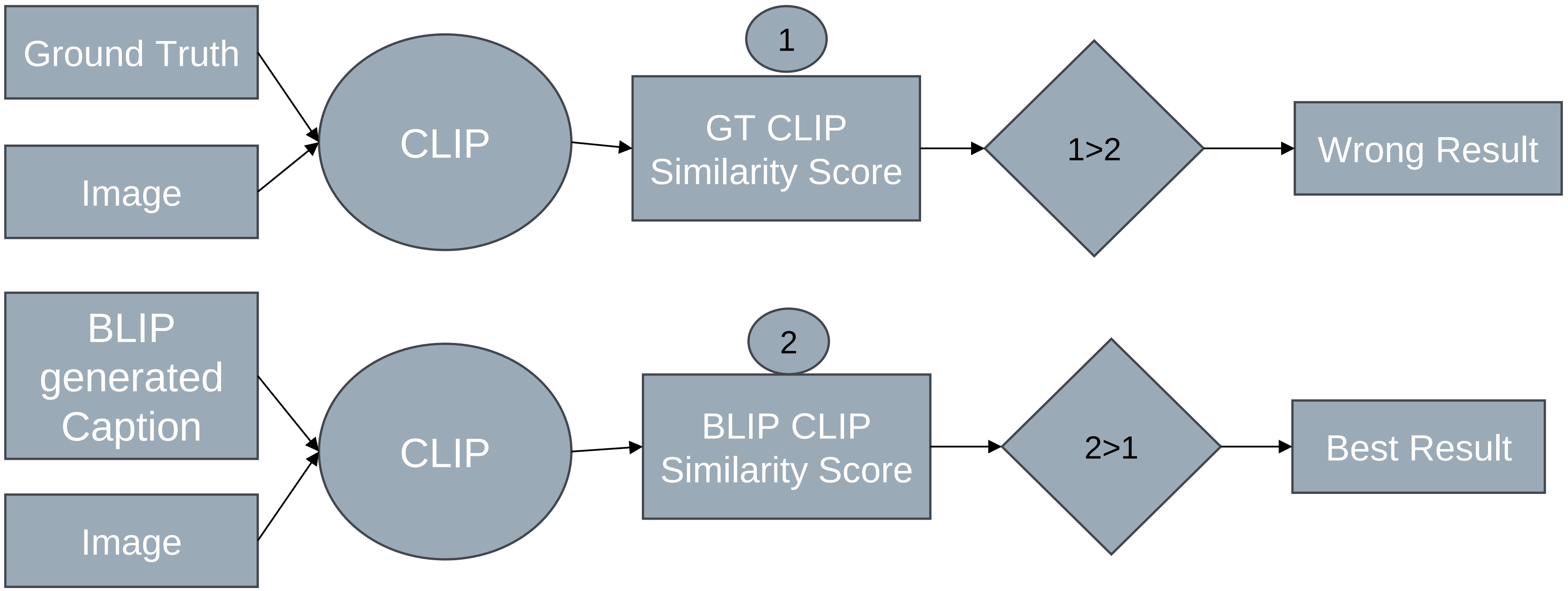
- To measure the efficacy of our fine tuning process, we proposed the following metrics for evaluation of generated prompts with ground-truth prompts.
  - **Number of samples:**

    200

Metrics	Best Case	Worst Case	Average
BLEU SCORE	1.0	0.3554	0.77
ROUGE 2	1.0	0.444	0.852
ROUGE L	1.0	0.6206	0.9204

Correct and Wrong Predictions

- We have done evaluation using using, where we have passed our image and ground truth and generated captions to our CLIP model and extract best and wrong predictions and displayed them in below slides,



Results

Correct Results

Results

Correct Results

Results

Wrong Predictions

Results

Wrong

Comparison

Ground Truth: Canvas_Shoes.Sneakers,and.Athletic.Shoes

Troika Predicted: Canvas_Shoes.Sneakers,and.Athletic.Shoes

Blip Predicted: Canvas_Shoes.Sneakers,and.Athletic.Shoes with black color and rubber Material.

Comparison

Ground Truth: Canvas_Shoes.Sneakers,and.Athletic.Shoes

Troika Predicted: Canvas_Shoes.Sneakers,and.Athletic.Shoes

Blip Predicted: Canvas_Shoes.Sneakers,and.Athletic.Shoes with unspecified color and rubber Material.

Comparison

Ground Truth: Canvas_Boots.Ankle

Troika Predicted: Canvas_Boots.Ankle

Blip Predicted: Canvas_Shoes.Sneakers,and.Athletic.Shoes with red color and rubber Material.

Conclusion

- The creation of enhanced prompt dataset using **CLIP embeddings** resulted in better description of images compared to original.
- The **fine-tuning** of Blip-2 LLM was implemented using LoRA and PEFT methods.
- The enhanced prompts were generated successfully for new input images and their evaluation was performed using various text & NLP evaluation metrics like **BLEU**, **ROUGE** and **semantic-similarity**.
- **Further optimizations** can include:
  - Inclusion of more object items e.g. mit-states or cifar datasets.
  - The embedding from CLIP can be generalized to include more objects for prompt enhancement provided sufficient fine-tuned data is available.

References

- Li, J., Li, D., Savarese, S. and Hoi, S., 2023, July. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (pp. 19730-19742). PMLR.
- Huang, S., Gong, B., Feng, Y., Zhang, M., Lv, Y. and Wang, D., 2024. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24005-24014).
- Lu, X., Guo, S., Liu, Z. and Guo, J., 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23560-23569).
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
- https://vision.cs.utexas.edu/projects/finegrained/utzap50k/

