Improving Visual Commonsense in Language Models via Multiple Image Generation

Reimplementation as part of an academic project at Sorbonne Université, ISIR

Saliou Barry & Carine Moubarak (Sorbonne Université, ISIR)

Can we enhance the visual commonsense of language models? Yes! By incorporating multiple image-based predictions into pre-trained LLMs, we significantly improve their performance on visual commonsense reasoning tasks.

Introduction

- **Context**: Large Language Models (LLMs) have shown impressive performance in text-based reasoning tasks but lack the ability to incorporate visual information.
- **Problem**:
  - Vision-Language Models (VLMs) are effective for visual tasks but struggle with commonsense reasoning that does not require visual input.
  - LLMs trained on textual data alone lack multimodal knowledge, limiting their ability to make visually grounded decisions.
- **Objective**: Develop an approach that enhances the visual commonsense reasoning of LLMs while maintaining their strong textual reasoning capabilities.

DataSets

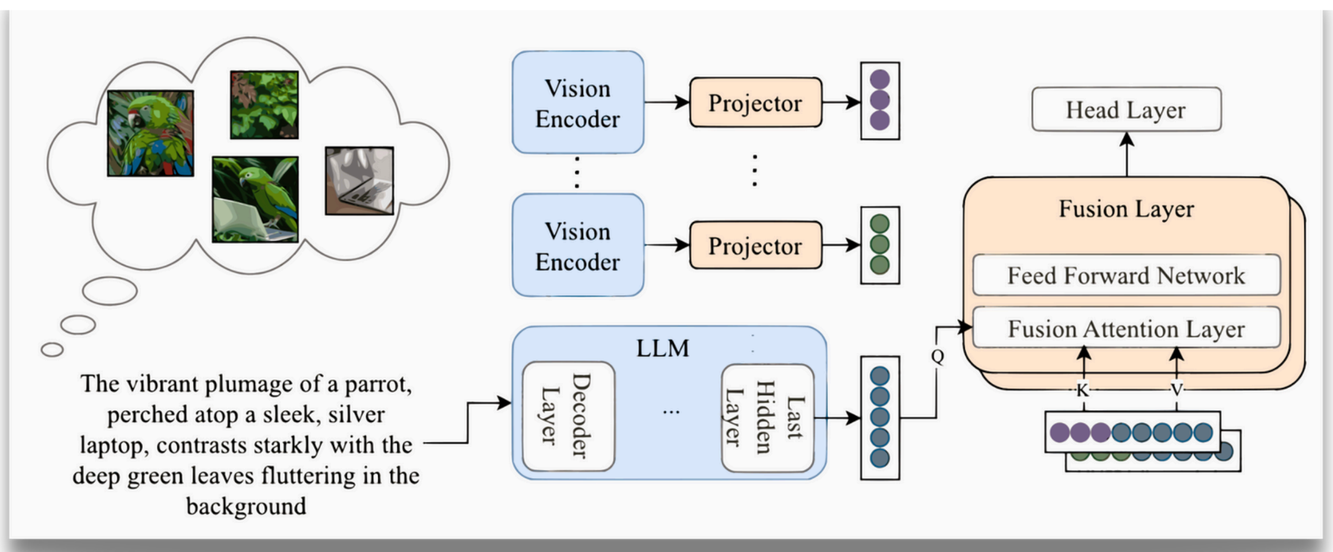
We use the **laion/220k-GPT4Vision-captions-from-LIVIS** dataset, a large-scale multimodal dataset containing image-text pairs sourced from the web. This dataset provides diverse and high-quality visual-text associations, essential for training our model on visual commonsense reasoning.

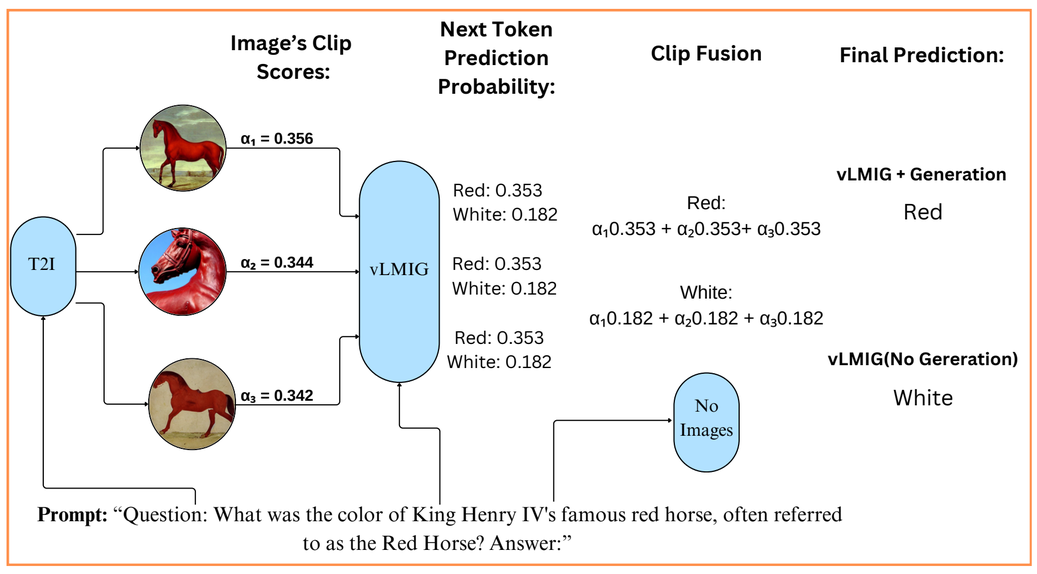
Training Process

Our model is trained on a real image-text dataset, where each image has a descriptive annotation. Unlike synthetic approaches, we directly integrate real multimodal associations for visual commonsense reasoning.

Key Components

- **Vision Encoder** : Extracts visual features.
- **Visual Token Projector (VTP)**: Converts visual features into text-compatible embeddings.
- **Custom LLM (GPT-initialized)**: Handles textual reasoning, fine-tuned for multimodal understanding.
- **Late Fusion Attention Layer (LFAL)**: Merges text and visual embeddings using attention mechanisms, ensuring selective visual integration.





Inference

At inference, the model generates multiple images from the input text and extracts features using a Vision Encoder. These are converted into pseudo-text tokens (VTP) and fused with textual embeddings via a Late Fusion Attention Layer (LFAL).
The final prediction is computed by aggregating probabilities from all generated images:

$$\sum_{i=1}^{k} P_\theta(x_t|x_1, \ldots, x_{t-1}, v_i)$$

An alignment score (CLIP-based) weighs the relevance of each image before final decision:

$$\sum_{i=1}^{k} f(\bar{x}_i, v_i) P(x_t|x_1, \ldots, x_{t-1}, v_i) + (1 - f(\bar{x}_i, v_i)) P(x_t|x_1, \ldots, x_{t-1})$$

This ensures robust decision-making by dynamically balancing text and vision.

Results

This section evaluates our method against baselines on visual commonsense and textual reasoning tasks, showcasing the benefits of our multimodal fusion approach.

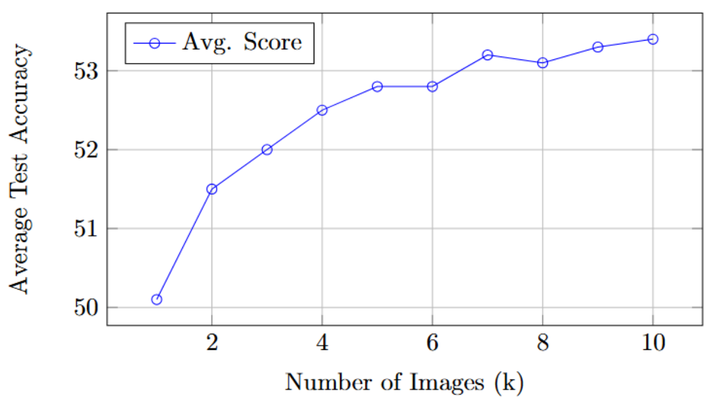
Evaluation on Various Tasks

Table 1 evaluates our model on a broader range of reasoning tasks: **visual commonsense** (**Memory Color, Color Terms, Object Shape, Relative Size**), **Commonsense Reasoning** (**BoolQ**), and **reading comprehension** (**SQuAD**). Results show that vLMIG improves both visual and textual understanding.

Model	Base Model	Visual Commonsense (Accuracy)	Commonsense Reasoning (Accuracy)	Reading Comprehension (Exact Match)	Reading Comprehension (F1 Score)
GPT-2	-	30.3	46.1	30.5	-
vLMIG	GPT-2	**38.6**	46.7	**32.2**	-
vLMIG + Image Gen + Scores	GPT-2	36.2	**62.0**	0.0	**9.35**
vLMIG + Image Gen	GPT-2	27.1	38.0	0.0	**9.35**
vLMIG (No Generation)	GPT-2	34.5	53.0	0.0	**9.35**

Impact of the number of generated images

Figure 2 illustrates the impact of the number of generated images per inference on performance. The curve shows a steady improvement in accuracy as k increases.
(**color , BoolQ**)



Conclusion

Strengths:

- Integrates visual and textual reasoning effectively.
- Improves performance in visual and textual benchmarks.

Limitations:

- Sensitive to the quality of generated images, which can mislead predictions.
- Errors may occur if irrelevant visual data influences decisions.
- High computational cost from generating multiple images.

Improvements:

- Enhance the quality and relevance of generated images.
- Explore VLMIG on complex visual tasks, like object relationships.

Reference

https://openreview.net/pdf?id=QP3EvD1AVa