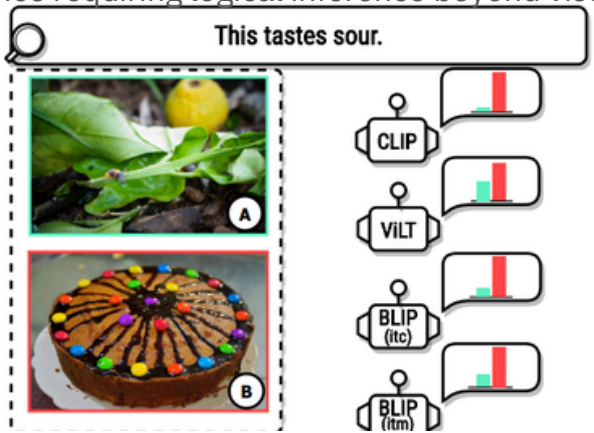
Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles

What is the problem?

- VL models, like CLIP and ViLT, struggle with commonsense reasoning, which is crucial for tasks requiring implicit understanding (e.g., identifying that "sour" refers to a lemon and not a cake).
- Existing VL datasets (e.g., COCO, CC 12M) focus more on object descriptions (nouns and adjectives) but lack verbs, particles, and relationships necessary for commonsense knowledge.
- This gap in commonsense limits the ability of VL models to handle real-world scenarios requiring logical inference beyond visual recognition.



What has been done earlier?

- Earlier approaches like CLIP and ViLT focused on vision-language alignment through contrastive or matching supervision, but they did not aim to integrate or enhance commonsense reasoning.
- Some models attempted to inject commonsense by leveraging caption datasets, but these approaches are limited by the corpus and statistical correlations, which fail to capture deeper reasoning or generalization.
- Research has explored other commonsense-focused tasks (e.g., Visual Question Answering), but these are not well suited for all types of VL models, especially those designed for image-text retrieval.

Sushobhan Tripathy,B421058

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

- Incorporating structured commonsense knowledge into image-text pairs without needing complex architectures, while avoiding domain shift at inference time, where external knowledge sources may be unavailable.
- The authors propose DANCE, a method that transforms knowledge graph entries into natural language riddles, allowing seamless integration into VL datasets. It conceals entity names using pronouns (e.g., “this item”), enabling models to learn relationships during training without requiring external knowledge at inference.
- This scalable approach enhances commonsense reasoning without compromising performance on standard tasks and introduces a retrieval-based benchmark to evaluate VL models' commonsense capabilities effectively.

Sushobhan Tripathy ,B421058