Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles

# Summary

The study aims to improve gaps in reasoning abilities and commonsense knowledge in recent popular vision-language (VL) models. Although these models are successful, they often lack an understanding of commonsense knowledge, which is considered to be essential for developing artificial general intelligence. With the goal to improve the commonsense ability of VL models, this research analyzes the causes of this constraint and proposes a scalable method called ”Data Augmentation with Knowledge graph linearization for Commonsense capability” (DANCE).

# Approach

The primary focus is on constantly augmenting existing datasets during training, instead of creating new training datasets. By using commonsense knowledge graphs such as ConceptNet, DANCE applies bidirectional sub-graph sequentialization in training to generate distinct description variations in VL datasets, which increases commonsense ability significantly. Also, the authors propose an innovative retrieval-based commonsense diagnostic benchmark that provides a methodical framework to evaluate DANCE’s efficiency. Through a focus on scalability, the paper’s methodology gives data augmentation preference over the collection of new datasets, guaranteeing useful applicability for a wide range of VL models.

# Validation

Through a multidimensional validation method that includes specific datasets, a retrieval-based commonsense diagnostic benchmark, and extensive tests on representative vision-language (VL) models, the authors validate their proposed DANCE technique. They present the first retrieval-based commonsense diagnostic benchmark, offering a unique framework for methodically evaluating the DANCE technique’s efficiency and VL models’ capacity for commonsense. Comprehensive tests on representative models show a notable improvement in commonsense abilities. Most importantly, the results show the scalability and effectiveness of the DANCE approach by showing improvement.

1

Benchmark performance provides additional evidence of DANCE’s efficiency and an in-depth evaluation of its impact on commonsense reasoning in VL models.

# Advantages

By addressing the crucial limitations, the DANCE technique dramatically improves the commonsense capabilities of vision-language models. By utilizing commonsense knowledge graphs like ConceptNet, it enhances generalization by allowing models to reason about a wider variety of concepts. With bidirectional sub-graph sequentialization, DANCE offers an innovative and scalable approach that eliminates the requirement for completely new datasets while enhancing models with a variety of commonsense knowledge.

# Disadvantages

The limitations of DANCE must be carefully considered, as its usefulness may differ based on the situation. The model’s capacity to learn accurate commonsense knowledge may be impacted by dependence on the biases and quality of the selected knowledge graph, such as ConceptNet. The introduction of computing complexity during training may provide difficulties, requiring a fair evaluation of computational efficiency compared to performance benefits.

# Future work

Recognizing the flaws in the commonsense abilities of current vision-language models (VL-models), the paper suggests the DANCE technique to include commonsense capability into VL-models. Even though DANCE exhibits significant improvements, the study notes that more research is required to meet gaps in knowledge. The paper specifically recommends concentrating on improving reasoning features in VL-models, such as mathematical and physical computations in practical situations. Future study on VL-models could potentially advance their capabilities by investigating and improving these reasoning elements.