THE NEURO-SYMBOLIC CONCEPT LEARNER: INTERPRETING SCENES, WORDS, AND SENTENCES FROM NATURAL SUPERVISION

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

1 Summary

The authors propose the Neuro-Symbolic Concept Learner (NS-CL), a model that learns visual concepts, words, and semantic parsing of sentences without explicit supervision instead learns by simply looking at images and reading paired questions and answers.Analogical to human concept learning, the perception module learns visual concepts based on the language description of the object being referred to. Meanwhile, the learned visual concepts facilitate learning new words and parsing new sentences.

2 Strengths of the proposal

1. They propose to use neural symbolic reasoning as a bridge to jointly learn visual concepts, words, and semantic parsing of sentences.
2. Introduces learning interpretable and disentangled representations for visual scenes using neural networks for joint reasoning with language modeling. Model includes Visual perception, Concept quantization, DSL and semantic parsing modules. Since visual representations and the concept representations has a fully differentiable design, supports gradient-based optimization during training.
3. NS-CL outperforms other methods in the task of learning visual concepts using a diagnostic question as well as in systematic study on visual features and data efficiency. The model uses no program annotations.

3 Weaknesses of the proposal

1. The visual task is simplistic and it is not obvious how this would generalize into other complex VQA tasks.
2. More experiments will draw strong conclusion considering CLEVR as a toy-dataset, even though it is suited for learning relational concepts.

4 Results

In the proposed model, visual perception module detects objects in the scene and extracts a deep, latent representation for each of them. The semantic parsing module translates an input question in natural language into an executable program given a domain specific language (DSL). The generated programs have a hierarchical structure of symbolic, functional modules, each fulfilling a specific operation over the scene representation. The model outperforms many state of the art models.

5 Discussion

Beyond static spatial relations and attributes, the model will not be able to perform for events such as for actions and interactions given that the semantic representations are yet to discover.