Project Proposal

EE 228 - Introduction to Deep Learning

Umit Yigit Basaran - 862393049
Biqian Cheng - 862135825
Emrullah Ildiz - 862393412

30 April 2023

The article "Neural Motifs: Scene Graph Parsing with Global Context" by Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi presents a novel approach to scene graph parsing, which is a computer vision task that involves generating a structured representation of objects and their relationships in an image. The authors propose a two-stage framework that first detects "neural motifs" and then uses them to guide object and relationship detection.

In the first stage of the framework, the authors use a graph neural network to detect common visual patterns across scenes. The neural motif detection stage involves two components: motif proposal and motif classification. The motif proposal component generates candidate motifs, while the motif classification component determines which candidates are actually motifs. The authors show that their approach outperforms previous methods that rely solely on object detection and relationship detection. The detected motifs capture important contextual information that can be used to improve object and relationship detection. In the second stage of the framework, the authors use the detected neural motifs to guide object and relationship detection. They propose a new loss function that encourages the model to predict objects and relationships that are consistent with the detected motifs. The loss function includes two components: a motif consistency loss and a standard detection loss. The motif consistency loss ensures that the predicted objects and relationships are consistent with the detected motifs, while the standard detection loss encourages accurate object and relationship detection.

The authors evaluate their approach on three benchmark datasets: Visual Genome, COCO, and VRD. Their approach outperforms previous state-of-the-art methods by a large margin on all three datasets. They also demonstrate that their model is able to generalize to scenes that contain previously unseen objects and relationships. The authors also propose a new evaluation metric, called "motif accuracy," which takes into account the accuracy of the predicted motifs. They argue that this metric is important because it encourages the model to focus on predicting motifs that are both frequent and semantically meaningful.

One potential limitation of their approach is that the resulting scene graphs may not be as comprehensive and informative as they could be. The scene graphs generated by the Neural Motifs model typically consist of object nodes and relationship nodes, where each relationship node represents a pre-defined motif. While this approach is effective for capturing common patterns in scene graphs, it may not be sufficient for more complex or diverse scenes. To achieve a better understanding of images and relation graphs, more structured, advanced, and informative relation graphs are needed. These relation graphs could include more detailed and nuanced relationships between objects, such as causal relationships or spatial relationships that are not captured by the pre-defined motifs used in the Neural Motifs model.

Moreover, the resulting scene graphs may not be easily interpretable or explainable, which can limit their usefulness in certain applications. More advanced graph visualization techniques and methods for interpreting the resulting scene graphs may be necessary to fully realize the potential of this approach. In summary, while the Neural Motifs approach is a promising step forward in scene graph generation, there is still room for improvement in the comprehensiveness and interpretability of the resulting scene graphs.

As a conclusion the neural motif networks are a new approach to structured graph representations of regularly appearing substructures in scene graphs. Although they show promise in improving prediction accuracy for relation labels based on object labels, they also raise concerns about their effectiveness in handling more complex or diverse scenes. Additionally, resulting scene graphs may lack interpretability or explainability, limiting their usefulness in some applications. In our project, to address these issues, we propose conducting a literature review and developing a new approach using other improvements such as inferring the bias on the motif model by using counterfactual causality, improving the encoder and decoder architectures and advanced dataset sampling techniques.