NeurIPS 2024 Paper Selection

CS 5588 – Advanced Topics in Generative AI

Bug Killers

RG-SAN: Rule-Guided Spatial Awareness Network for End-to-End 3D Referring Expression Segmentation

Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, Rongrong Ji

- Paper Link: https://openreview.net/forum?id=r5spnrY6H3
- GitHub Repository: https://github.com/sosppxo/RG-SAN
- Justification for Selection: The paper presents the RG-SAN model, which significantly enhances 3D Referring Expression Segmentation (3D-RES) by effectively modeling spatial relationships among entities described in natural language, addressing limitations of traditional methods that often lead to over-segmentation and mis-segmentation due to insufficient spatial awareness.

RG-SAN: Rule-Guided Spatial Awareness Network for End-to-End 3D Referring Expression Segmentation

Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, Rongrong Ji

Summary of Key Contributions:

- **Introduction of RG-SAN:** The authors propose RG-SAN as a novel approach that enhances the modeling of spatial relationships among all entities in referring expressions, significantly improving the model's ability to accurately segment 3D objects based on textual descriptions.

- **Text-driven Localization Module (TLM):** The TLM is designed to precisely localize all instances mentioned in the referring expressions by iteratively refining their positional information. This module enhances the model's performance by effectively leveraging spatial information, which is crucial for accurate segmentation in complex scenes.

- **Rule-guided Weak Supervision (RWS):** The RWS strategy utilizes dependency tree rules to guide the positioning of core instances based solely on the target instance's location. This innovative approach allows the model to effectively handle the challenges of weak supervision, leading to improved accuracy in identifying and segmenting the target objects.

E2E-MFD: Towards End-to-End Synchronous Multimodal Fusion Detection

Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, DaixunLi, Wenbo Huang, Yunsong Li, Xue Yang

- Paper Link: https://openreview.net/forum?id=47loYmzxep

- GitHub Repository: https://github.com/icey-zhang/E2E-MFD

- Justification for Selection: The paper presents a significant advancement in the field of multimodal image fusion and object detection, particularly relevant for applications in autonomous driving and remote sensing. The proposed E2E-MFD framework integrates image fusion and object detection into a single training phase, enhancing efficiency and performance compared to traditional methods that often rely on complex, multi-step processes.

E2E-MFD: Towards End-to-End Synchronous Multimodal Fusion Detection

Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, DaixunLi, Wenbo Huang, Yunsong Li, Xue Yang

Summary of Key Contributions:

- **End-to-End Framework:** E2E-MFD introduces a pioneering approach that integrates image fusion and object detection into a single-stage, end-to-end framework. This design significantly enhances the outcomes of both tasks by streamlining the training process, which traditionally involved complex multi-step methods.

- **Gradient Matrix Task-Alignment (GMTA):** The authors propose a novel GMTA technique that optimizes the training process by evaluating and quantifying the impacts of the image fusion and object detection tasks. This method aids in stabilizing the training process and ensures convergence to an optimal configuration of fusion detection weights, thereby minimizing the adverse effects typically associated with multi-task learning.

- **Object-Region-Pixel Phylogenetic Tree (ORPPT):** The introduction of the ORPPT allows for the extraction of features at multiple granularities, facilitating the interaction between the fusion and detection tasks. This innovative structure enhances the model's ability to reconcile fine-grained details with semantic information, ultimately improving the performance of both image fusion and object detection.

You Only Cache Once: Decoder-Decoder Architectures for Language Models

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, Furu Wei

- Paper Link: https://openreview.net/forum?id=25Ioxw576r
- GitHub Repository: No GitHub, but they publish at https://aka.ms/GeneralAI
- Justification for Selection: The paper introduces YOCO, a novel decoder-decoder architecture for large language models that significantly enhances inference efficiency while maintaining competitive performance. By caching key-value pairs only once, YOCO reduces GPU memory consumption by approximately 80x for large models, making it feasible to deploy long-context language models on standard hardware. The authors provide detailed hyperparameters and experimental setups, ensuring that the results can be reproduced effectively, as indicated in the appendix.

E2E-MFD: Towards End-to-End Synchronous Multimodal Fusion Detection

Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, DaixunLi, Wenbo Huang, Yunsong Li, Xue Yang

Summary of Key Contributions:

- **Memory Efficiency:** YOCO significantly reduces GPU memory consumption by caching key-value pairs only once, which allows for a reduction of approximately 80x in memory usage for large models compared to traditional Transformers. For instance, a 65B model can operate with much lower memory requirements, making it feasible to deploy long-context language models on standard hardware.

- **Improved Prefilling Latency:** The architecture enables early exits during the prefill stage, which dramatically speeds up the process. YOCO reduces the prefilling latency for a 512K context from 180 seconds to less than 6 seconds, showcasing a speedup of 71.8x for long sequences.

- **Scalability and Performance:** YOCO demonstrates competitive performance across various tasks and scales effectively with increased training tokens and context lengths, achieving near-perfect accuracy in long-context retrieval tasks. The model can handle up to 1 million tokens while maintaining high accuracy, positioning it as a strong candidate for future large language models.

Individual Contributions

Team Member Name	Contributions
Hui Jin	Prepared class slides and final submission
Jayadithya Nalajala	Identified and summarized Paper 1
Saniya Pandita	Identified and summarized Paper 2
Sai Jahnavi Devabhakthuni	Identified and summarized Paper 3