Plano de Estudo para Dissertação

Exploring Position Trainable Transformers for Computer Vision

Universidade do Minho - Departamento de Matemática

Rui Morais - PG54403

Palavras-chave: Position Trainable Transformer, Computer Vision, Attention, Vision Transformer.

Introdução

Transformers [8] surgiu como uma abordagem promissora para tarefas de visão computacional, após ter sido inicialmente concebida para processamento de linguagem natural [5] . A utilização de um mecanismo de atenção permite que os Transformers aprendam relações contextuais complexas de forma eficiente. Em 2020, investigadores descobriram que o Transformer podia ser adaptado ao processamento de imagens, o que deu origem à arquitetura "Vision Transformer"(ViT) [3]. Diferente das redes convolucionais (CNN) [2][7], que dependem de operações de convolução locais, o Transformer consegue capturar dependências globais de maneira direta, tornando-se uma escolha atraente para tarefas complexas de visão computacional. O Transformer, aplicados no contexto de super-resolução (SR) [10][6][1][4], Segmentação [9], Detenção de Objetos [4], entre outros, pode demonstrar grande potencial no melhoramento da performance dos modelos baseados em CNNs.

Objetivos

- Desenvolver uma compreensão profunda dos Transformers aplicados à visão computacional.
- Dominar os fundamentos de Position Trainable Transformers (PTTs).
- Explorar e implementar Vision Transformers e Position Trainable Transformers.
- Experimentar com embeddings de posição treináveis e analisar seu impacto.
- Implementar aplicações de transformers, como DETR, CLIP, ALIGN entre outros.
- Realizar um projeto final de pesquisa que compare o desempenho dos PTTs com transformers convencionais.

Metodologia

O plano de estudos está organizado em módulos mensais com tarefas semanais, distribuídas da seguinte forma:

- Fundamentos de Transformers e Visão Computacional: Estudo inicial sobre o funcionamento dos Transformers e de CNNs aplicadas à visão computacional.
- Vision Transformers: Exploração detalhada dos Vision Transformers, incluindo treino e fine-tuning de modelos com diferentes datasets.

- Exploração de Position Trainable Transformers: Estudo e experimentação com embeddings de posição treináveis.
- Aplicações de Transformers na Visão Computacional: Implementação de aplicações práticas, como DETR e CLIP, para tarefas multimodais.
- Aprendizagem Auto-Supervisionada: Investigação de técnicas de aprendizagem auto-supervisionada aplicadas a transformers.

Resultados Esperados

Ao final do plano de estudos, espera-se que o estudante:

- Domine os conceitos fundamentais e as principais arquiteturas de transformers aplicadas à visão computacional.
- Seja capaz de implementar e ajustar ViTs e PTTs, compreendendo os impactos das configurações de embeddings de posição.
- Desenvolva um projeto de pesquisa, aplicando PTTs a uma tarefa específica, e compare seu desempenho com transformers convencionais.

Conclusão

Este plano de estudos oferece uma estrutura sólida para o desenvolvimento de competências avançadas em transformers para visão computacional, com foco em Position Trainable Transformers. Através de leituras guiadas, prática de implementação e um projeto de pesquisa, o estudante estará bem preparado para contribuir para o avanço da pesquisa e das aplicações de transformers na visão computacional.

Referências

[1] Neeraj Baghel, Shiv Ram Dubey, and Satish Kumar Singh. Srtransgan: Image super-resolution using transformer based generative adversarial network. *arXiv preprint arXiv:2312.01999*, 2023.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Xiaoqian Huang, Detian Huang, Qin Huang, Caixia Huang, Feiyang Chen, and Zhengjun Xu. Dtsr: detail-enhanced transformer for image super-resolution. *The Visual Computer*, pages 1–18, 2023.

[5] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

[6] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022.

[7] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[8] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[9] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[10] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408, 2016.