Computer Vision Project Proposal

Lucas PERRIER, Tiago MACHADO MONTEIRO

Proposal:

The detection and monitoring of cracks in concrete structures are essential for ensuring structural safety and longevity. Conventional convolutional neural networks (CNNs) have achieved high accuracy in automated crack detection but lack interpretability, limiting their acceptance in real-world civil engineering applications. This project proposes an explainable Vision Transformer (ViT) framework for crack detection in construction materials. By integrating attention-based architectures with explainability tools such as Grad-CAM and SHAP, the project aims to achieve both high predictive accuracy and human-interpretable decision maps. The study will benchmark ViT against CNN baselines under varied environmental conditions, such as illumination and surface texture, using publicly available datasets (SDNET2018 and CCIC). The final outcome will include an interpretable, robust crack detection model and a visualization dashboard for engineering decision support.

The dataset used for training and evaluation will consist of the SDNET2018 and Concrete Crack Image Datasets, and we will preprocess the dataset by applying normalization, resizing and augmentation (rotation, lighting, and gaussian noise). We will train a baseline CNN type model (ResNet 50) and finetune a pre-trained ViT model using transfer learning on image net weights, on the same dataset. We will integrate explainability AI tools, namely Grad-CAM and SHAP, to visualize model attention and decision regions. We will evaluate interpretability with localization accuracy and faithfulness metrics; model performance using accuracy, f1-score, and AUC; and model robustness by comparing both models under new lighting and material samples.

We expect to produce a trained and interpretable Vision transformer capable of detecting and localizing cracks with more than 95 percent accuracy, show model focus regions using heatmaps and attention maps. Carry out a comparative analysis of ViT's and CNN's advantages in robustness and interpretability and showcase results in a report with visualizations.

We have assembled a list of key references which include comprehensive reviews on crack detection and more general method papers which we have used for this project proposal and will use to deepen our understanding of using explainable vision transformers for crack detection.

Bibliography for literature review:

Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR 2021. https://arxiv.org/abs/2010.11929

Golding, V.P.; Gharineiat, Z.; Munawar, H.S.; Ullah, F. *Crack Detection in Concrete Structures Using Deep Learning.* Sustainability 2022, 14, 8117. https://doi.org/10.3390/su14138117

Kashefi, R.; Barekatain, L.; Sabokrou, M.; Aghaeipoor, F. *Explainability of Vision Transformers: A Comprehensive Review and New Perspectives.* arXiv 2023. https://arxiv.org/abs/2311.06786

Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D*. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. International Journal of Computer Vision 2020, 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7

Zhang, X.; Wang, H.; Hsieh, Y.-A.; Yang, Z.; Yezzi, A.; Tsai, Y.-C. *Deep Learning for Crack Detection: A Review of Learning Paradigms, Generalizability, and Datasets.* arXiv 2025. https://arxiv.org/abs/2508.10256