

Pavement Distress with VMamba

Intro

With the importance of automobiles in our daily lives, catching distressed pavement in a timely manner is crucial for ensuring the safety of our transportation infrastructure. Researchers have implemented machine learning algorithms to classify cracked pavement as a means to create an early warning system for distressed pavement. Various models such as the Pict[1], Swin-S[2], and EfficientNet-B3[3], showed strong classification skills and have laid the foundation for future models to surpass them.

Pavement distress classification methodology

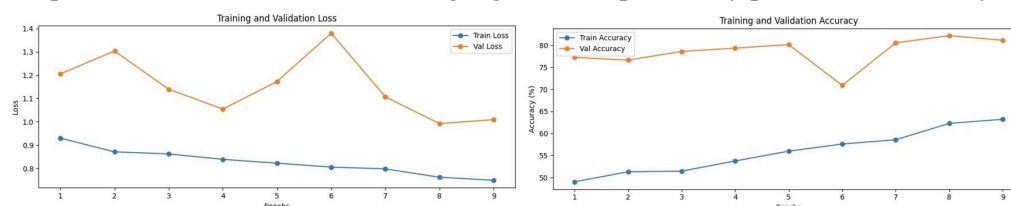
The PicT model has shown great accuracy with pavement distress classification, so this section will cover the methodology of the model to better understand how the process of image classification is conducted. In order to classify the type of distress present in a segment of pavement, the model must first identify if the image has any distress present. To do this the model utilizes distillation learning[4] in their “Patch Labeling Teacher”[1] which enhances the exploration of discriminative information in patches by employing a teacher-student learning scheme. Essentially, a larger model, the teacher model, can do a much more detailed analysis of the image and then pass on its predictions to the student model. Thus, giving the student model the power of the larger model with far less computations. Then clusters of images are grouped together in patches based on the model's risk prediction for each image. During training the Swin Transformer[2] blocks are utilized to extract visual features referred to as tokens[5]. The model calculates the loss for both the images and patches to more effectively learn from the different levels of data. The result was an astounding 92.2% Top-1 accuracy score[1], a truly remarkable achievement for pavement distress models.

VMamba is pavement distress.

Vision Mamba or VMamba for short is a modern machine learning architecture built to efficiently process images in a revolutionary way[10]. By partitioning its Mamba attention block over the whole picture, the model can use 74.5%[7] fewer parameters than the Transformers model[6] while having nearly identical performance. This advancement allows for image classification models to use high-resolution images which makes pavement distress models much more accurate. The advantages of VMamba lend themselves to other applications, such as video, point cloud, and multi-modal data[8].

My Code and results

I was assigned to emulate the distillation learning feature that VMamba and other architectures use. I decided to use EfficientNet-B7[3] as my teacher model due to it being accurate and large and ResNet-50[9] as the student model. I configured the model to take in a pavement dataset from ImageNet[6] called pavement dataset 2. From there I constructed the distillation parameters as well as the distillation loss function. After training the model on 10 epochs, it achieved a top-1 accuracy of nearly 82%. Considering the simplicity of the model, this result is a success and shows that distillation learning is a powerful tool for machine learning algorithms, specifically pavement distress systems.



References

- [1] Zhang, Y., Wang, Y., Liu, Y., & Zhang, J. (2022). PicT: A slim weakly supervised vision transformer for pavement distress classification. In *Proceedings of the ACM Multimedia Conference* (pp. 1-9). ACM.
- [2] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *arXiv*. <https://doi.org/10.48550/arXiv.2103.14030>
- [3] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv*. <https://doi.org/10.48550/arXiv.1905.11946>
- [4] Wang, L., & Yoon, K.-J. (2022). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6), 3048–3068. <https://doi.org/10.1109/TPAMI.2021.3055564>
- [5] Ryoo, M. S., Piergiovanni, A. J., Tan, M., & Angelova, A. (2021). TokenLearner: What can 8 learned tokens do for images and videos? *arXiv*. <https://doi.org/10.48550/arXiv.2106.11297>
- [6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. <https://doi.org/10.1109/CVPR.2009.5206848>
- [7] Chen, Z., Shamsabadi, E. A., Jiang, S., Shen, L., & Dias-da-Costa, D. (2024). Vision Mamba-based autonomous crack segmentation on concrete, asphalt, and masonry surfaces. *arXiv*. <https://doi.org/10.48550/arXiv.2406.16518>
- [8] Zhao, M., Zhang, Y., Liu, X., & Zhang, L. (2024). A survey on Vision Mamba: Models, applications, and challenges. *arXiv*. <https://doi.org/10.48550/arXiv.2404.18861>
- [9] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv*. <https://doi.org/10.48550/arXiv.1512.03385>
- [10] Liu, X., Zhou, Y., & Guo, B. (2024). VMamba: Visual state space model. *arXiv*. <https://doi.org/10.48550/arXiv.2401.10166>