Hybrid CNN-Transformer Model for Cancer Detection in Histopathology Images

Yashraj Shanker

Carnegie Mellon University

Why Hybrid CNN-Transformer Models?

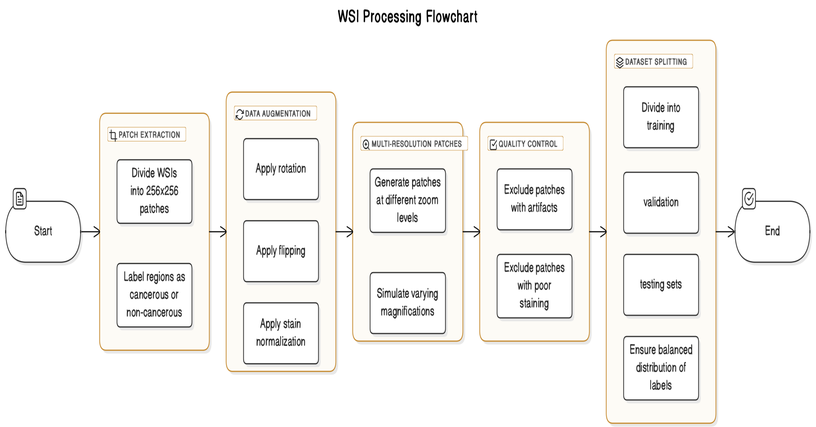
Cancer detection in histopathology requires analyzing high-resolution Whole Slide Images (WSIs), which pose challenges due to their large size, complex structures, and staining variability. Vision Transformers excel at providing contextual awareness but are computationally expensive and data-intensive. On the other hand, CNNs are effective at extracting local features such as textures and edges but struggle to capture broader contextual information needed for comprehensive tissue analysis.

- The Hybrid CNN-Transformer model combines the strengths of CNNs and Transformers.
- CNNs capture fine-grained local features like textures and edges.
- Transformers provide powerful attention mechanisms for global context.
- The synergy enables better hierarchical representation of tissue structures.
- The model enhances detection of cancerous regions in histopathology images.
- Improved performance is achieved in medical imaging tasks.
- The model overcomes limitations, enabling accurate and efficient cancer diagnosis.
- Provides a robust solution for large-scale histopathology data analysis.

Methods

Dataset Preparation

The dataset preparation process begins with extracting 256x256 pixel patches from Whole Slide Images (WSIs) using annotations to label regions as cancerous or non-cancerous. Data augmentation techniques, including rotation, flipping, and stain normalization, are applied to enhance robustness and account for variability in staining and tissue morphology. Multi-resolution patches are generated to simulate varying magnifications for hierarchical analysis, mimicking a pathologist's workflow. Quality control ensures only high-quality patches are included, while the dataset is split into training, validation, and testing sets with a balanced distribution of labels. This comprehensive process ensures the model can effectively learn from diverse and reliable data.



WSI Processing Flow Chart

Method

Architecture

The workflow begins with Whole Slide Images (WSIs) and their annotations as input data. Hierarchical patches, including cell-level, patch-level, and region-level, are extracted and structured into datasets during the data preparation stage. These patches are then encoded using a U-Net encoder to generate feature maps, which are processed through hierarchical transformers at the cell, patch, and region levels. The extracted features are aggregated using a feature fusion module to enable accurate classification. The model is optimized using loss functions, learning rate schedulers, and auxiliary outputs to ensure robust performance. Attention maps and heatmaps are generated to visualize cancerous regions effectively. Finally, the optimized model is scaled for deployment on GCP, with integration enabled through a REST API for seamless accessibility and scalability.

Result and Discussions

Key metrics such as accuracy, precision, recall, and F1 score are essential for evaluating the performance of a Hybrid CNN-Transformer Model for cancer detection in histopathology images. Accuracy measures overall classification performance, while precision ensures predicted cancerous regions are truly malignant. Recall assesses the model's ability to detect all cancerous regions, and the F1 score balances precision and recall. Together, these metrics validate the model's reliability and effectiveness in handling the complexity of histopathology data for clinical applications.

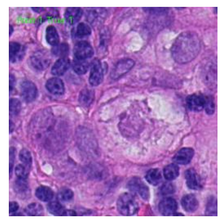
Training Averages

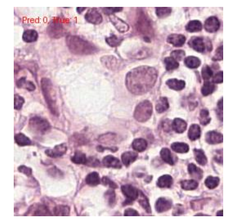
Metric	Description	Value
train_loss	Model error on training data	0.0849
train_accuracy	Correct predictions on training data	0.9727
val_loss	Model error on validation data	0.3389
val_accuracy	Correct predictions on validation data	0.9141
val_precision	Correct positive predictions	0.9453
val_recall	Detected actual positives	0.9208
val_f1_score	Precision-recall balance	0.9315
val_auc	Class separation ability	0.9528

Testing Results

Metric	Value
Accuracy	0.0935
Recall	0.0935
F1 Score	0.1711

The model's performance, with an accuracy and recall of 0.0935 and an F1 score of 0.1711, highlights the need for improvement. Enhancements could include advanced preprocessing (e.g., stain normalization), a larger, more diverse dataset, and fine-tuning hyperparameters like learning rate and transformer depth. Exploring alternative architectures like Swin Transformers, ensemble methods, or attention-based loss functions could also improve feature extraction and generalization, leading to better cancer detection accuracy.





References

Geert Litjens et al. (2016). "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Radboud University Medical Center. pp. 824-6525

Chen, R. J., et al. (2021). "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning." In: IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 123-132