WeCAViT: Weighted Ensemble of CNN, Attention, and a Visual Transformer - A Novel Ensemble Neural Network Architecture for Accurate Pneumonia Diagnosis From Chest X-Rays

Jumabekov, Arnur (School: Haileybury Almaty)

Pneumonia is a potentially lethal lung disease characterized by inflammation of the alveoli. Chest X-ray (CXR) imaging is the primary diagnostic tool, but human interpretation requires significant time, increasing the risk of misdiagnosis, especially in large-scale screenings. AI-based solutions using Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have been proposed to assist radiologists' decisions. However, CNNs often fail to capture broader context, while ViTs struggle with localized details and often need large training sets. This research proposes a WeCAViT model, aiming to leverage the strengths and mitigate the weaknesses of previous methods. Samples of pneumonia and healthy lungs were obtained from publicly available datasets, e.g., PneumoniaMNIST. The proposed model utilizes an ensemble architecture combining multiple CNN encoders, Spatial and Channel Attention Blocks, and a Vision Transformer. CNN encoders extract local features at varying scales (28x28, 56x56, 112x112), which are then decoded and refined via Spatial and Channel Attention mechanisms to highlight significant details. Feature maps are weighted via learnable parameters, merged, and processed using a ViT modeling long-range dependencies to detect subtle pathological signs. WeCAViT achieved 93.6% test accuracy, surpassing many state-of-the-art benchmarks (pure CNN/ViT, hybrid models) in specificity and overall robustness, even with limited training data. This study demonstrates the efficacy of the proposed architecture for automated pneumonia diagnosis, particularly relevant for analyzing large-scale data in mass population CXR screenings. Future research will evaluate the model's performance and generalizability on diverse datasets and enhance its interpretability.

Awards Won: