EGNAS: Neural Architecture Search for GNN Medical Imaging Models

Authors

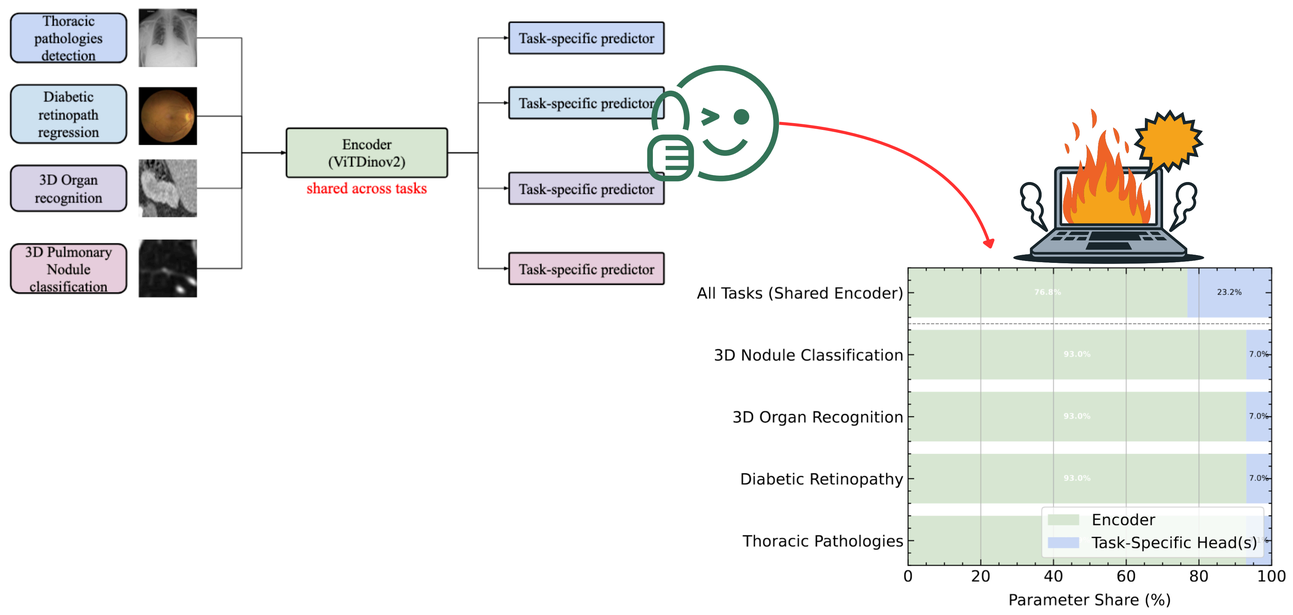
Hadjer Benmeziane*, Abderaouf Gacem[+], Kaoutar El Maghraoui‡, and Sarra Benmeziane[π]

Affiliations

*IBM Research Europe, 8803 Ruschlikon, Switzerland
‡IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[+] INSA Lyon
[π] EPS Mohamed Boudiaf Ouargla

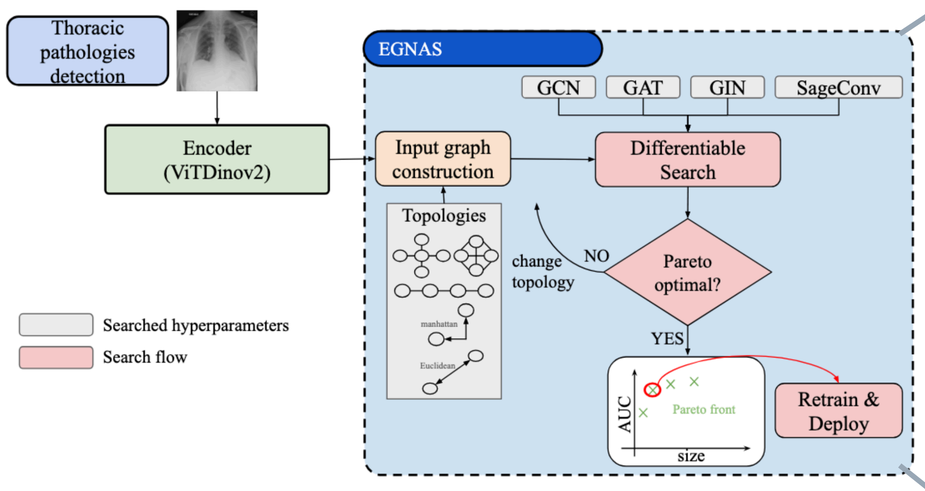
Context

- Deep learning has advanced medical imaging diagnosis but remains difficult to deploy in resource-constrained clinical environments.
- A common approach uses a shared encoder for multiple tasks to reduce redundancy, yet task-specific prediction heads remain memory-heavy.
- As the number of tasks grows, these heads collectively consume a significant portion of the model's size.
- Graph Neural Networks (GNNs) offer a more efficient and expressive alternative to standard MLP heads.



Can we design smaller and efficient heads for multiple medical tasks ?

Methodology



Search Space: EGNAS searches over GNN operator types (e.g., GCN, GAT, GIN), graph topologies (e.g., slice-based, similarity-based), and architectural hyperparameters (e.g., depth, hidden size, batch norm). Each architecture is represented as a differentiable graph with softmax-weighted operations.

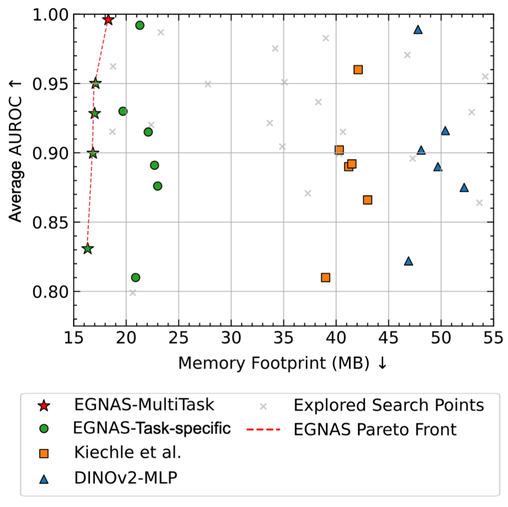
The loss function combines task-specific prediction loss with a memory cost term, balancing model size and inference efficiency. EGNAS maintains a Pareto front of non-dominated architectures to guide search toward optimal trade-offs.

EGNAS performs a two-phase search: first, it uses a differentiable search to optimize architectures jointly; then, it discretizes the top candidates for final evaluation. Gumbel-softmax enables smooth optimization over discrete design choices.

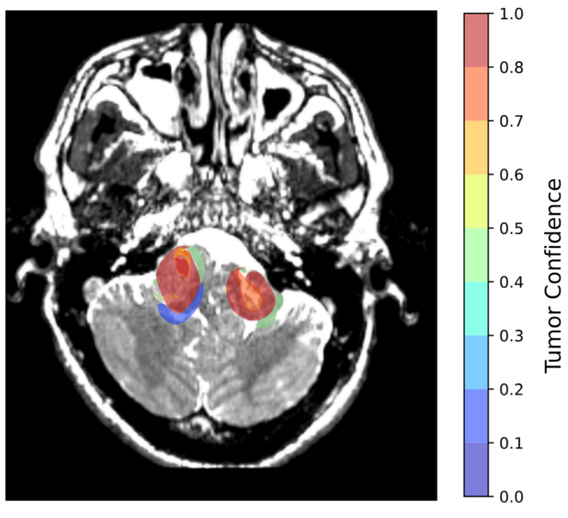
Evaluation Results

We evaluated EGNAS on six medical image classification tasks from the MedNIST3D dataset, using a shared encoder and task-specific heads. EGNAS was benchmarked against two strong baselines: a standard DINOv2-MLP setup and the GNN-based architecture from Kiechle et al. (2024).

- Across all tasks, EGNAS achieved state-of-the-art accuracy, while reducing task-head memory usage by an average of 2.1× compared to MLPs and 1.9× compared to previous GNN methods.
- It also provided faster inference, with average runtime reductions of 54% and 45% respectively. Importantly, EGNAS models consistently occupied or approached the Pareto front, demonstrating a superior trade-off between predictive performance and resource efficiency.



To validate its practical utility, we deployed EGNAS in a real-world clinical setting in Algeria. Operating on a low-spec dual-core Intel laptop with no GPU, EGNAS successfully detected brain tumors with 78% IoU in under 1.5 seconds per image, within a 300MB memory budget. These results confirm EGNAS's potential for accurate, fast, and efficient medical AI in low-resource environments.



Conclusion

We introduced EGNAS, a Pareto-efficient neural architecture search framework that discovers lightweight, task-specific GNN heads for multi-task medical imaging. Unlike prior work that optimizes solely for accuracy, EGNAS balances predictive performance with memory and runtime efficiency through a differentiable, graph-structured search.

References

Benmeziane, H. et al. (2024). Medical neural architecture search: Survey and taxonomy. IJCAI.

Yang, J. et al. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10(1):41.

Kiechle, J. et al. (2024). Graph neural networks: A suitable alternative to MLPs in latent 3D medical image classification? In International Workshop on Graphs in Biomedical Image Analysis.