Localization-Guided Supervision for Robust Medical Image Classification by Vision Transformers

Sagi Ben Itzhak[1], Nahum Kiryati[1], Orith Portnoy[2], Arnaldo Mayer[2]

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel

[2] Diagnostic Imaging Department at Sheba Medical Center Affiliated with the School of Medicine, Tel Aviv University, Tel Aviv, Israel

Introduction

Problem

Medical Image Analysis

- **Small datasets (annotation cost)**
- **Technical variability (different scanners & protocols)**

Small datasets + Technical variability → Overfitting → Poor Generalization & Reduced Performance

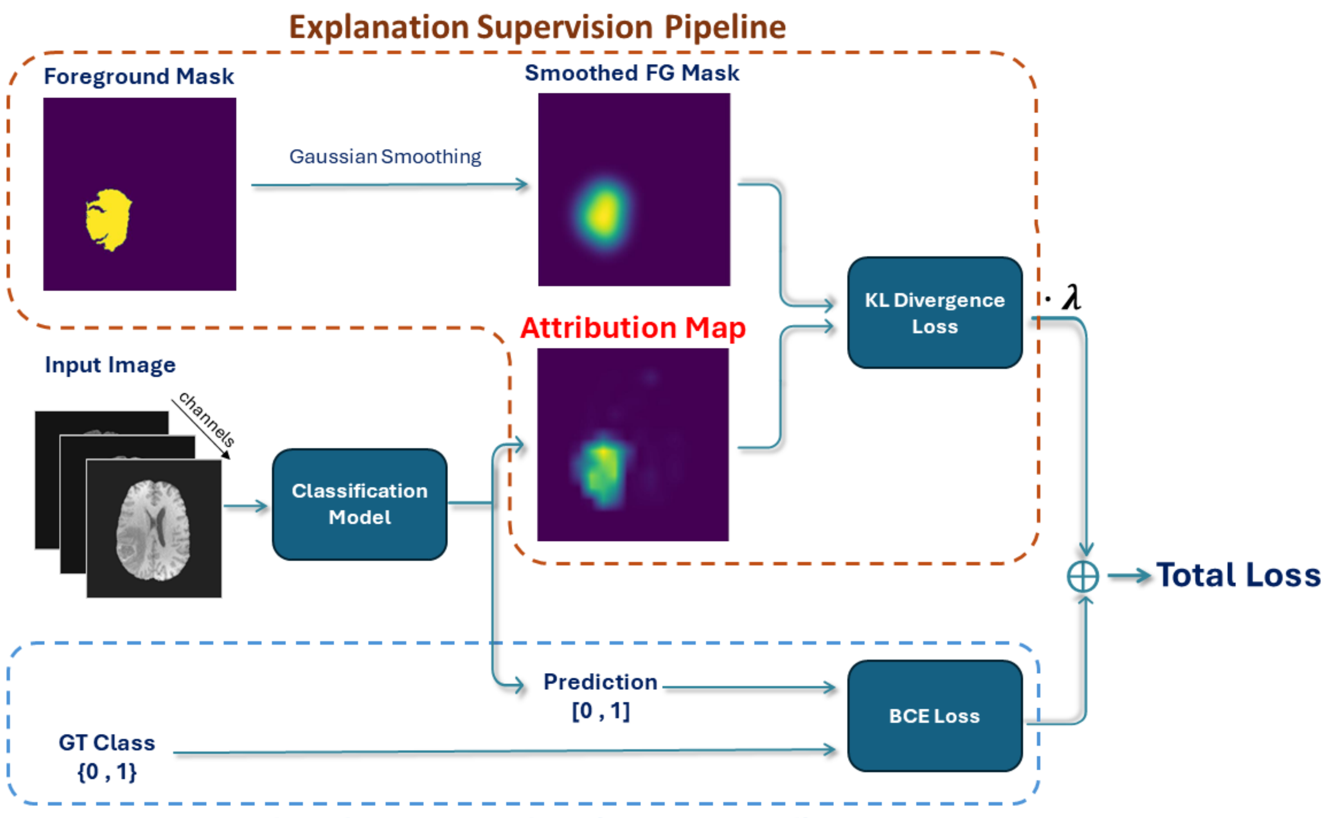
Suggested Approach

Robust image classification via **explanation supervision**.

LGM-ViT (Localization-Guided Medical Vision Transformer): End-to-end training of ViT-based classification models with explanation supervision.

Methods

Overview of LGM-ViT Training Framework

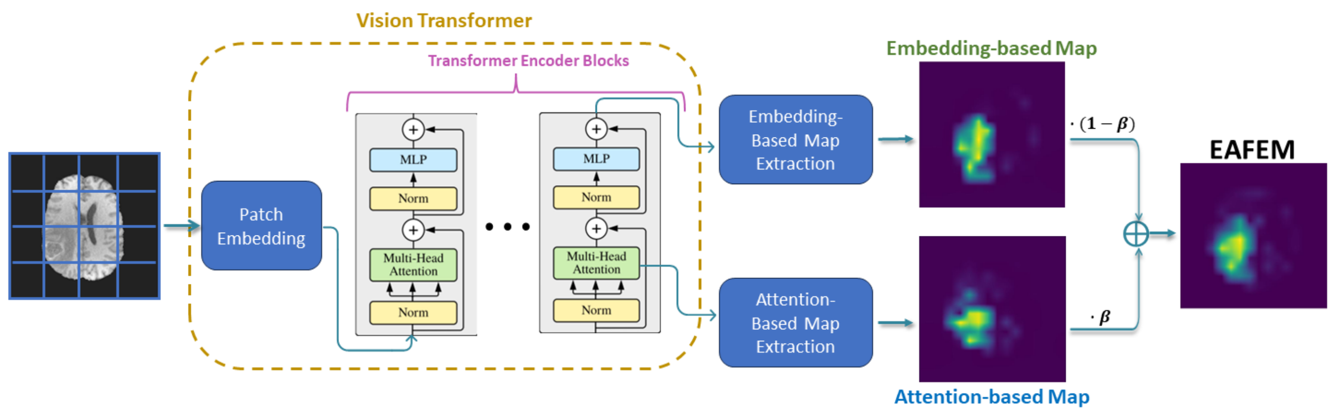


- **Classification Pipeline:** encourages correct predictions, improving accuracy.
- **Explanation Supervision Pipeline** encourages correct predictions **"for the right reasons"**, enhancing generalization and robustness.

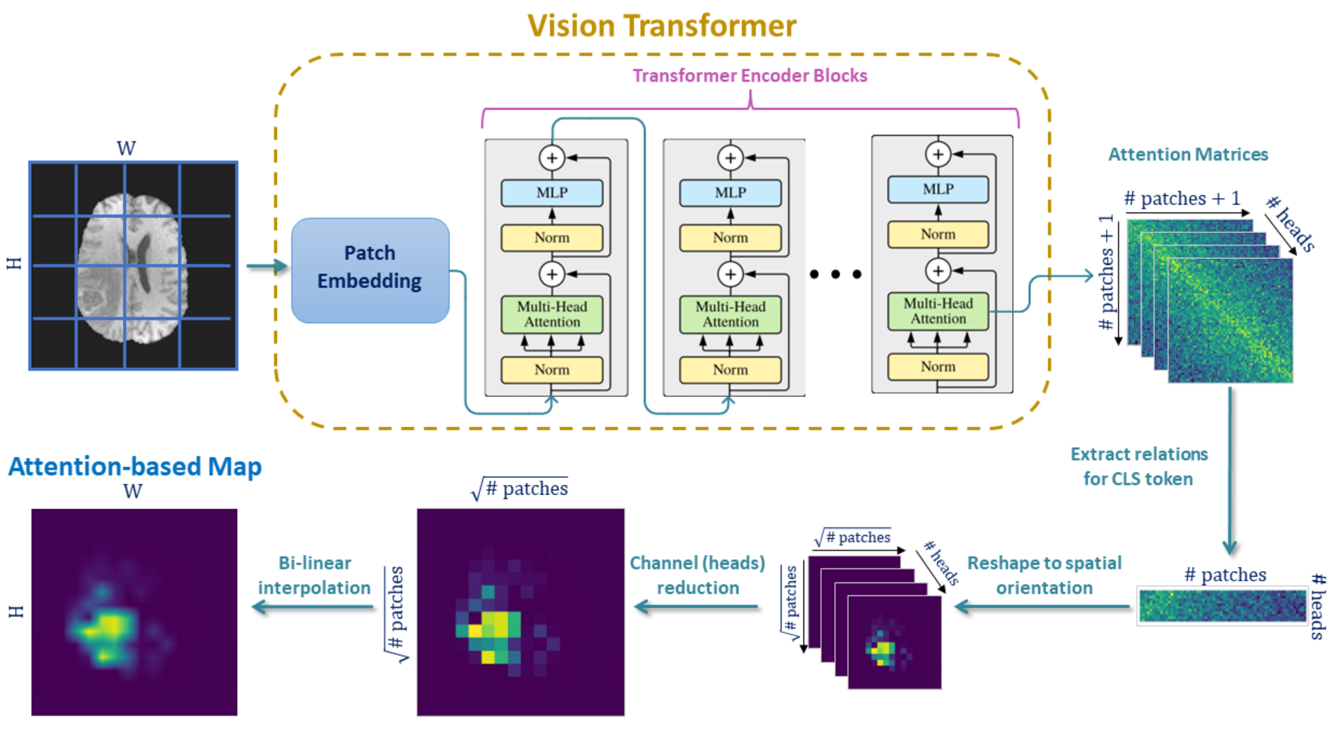
Attribution Map

EAFEM (Embedding-Attention Fused Explanation Map), used as Attribution Map in LGM-ViT, combines attention information with feature representation:

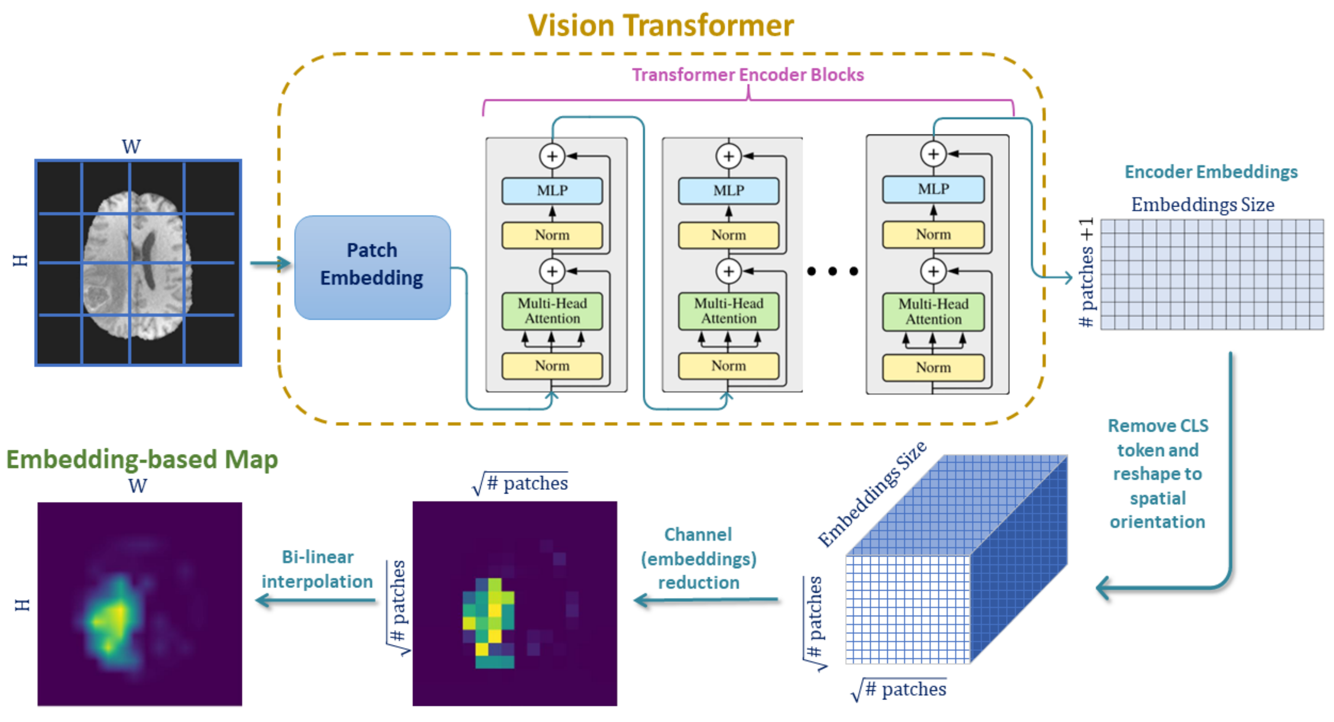
EAFEM



Attention-based Map



Embedding-based Map



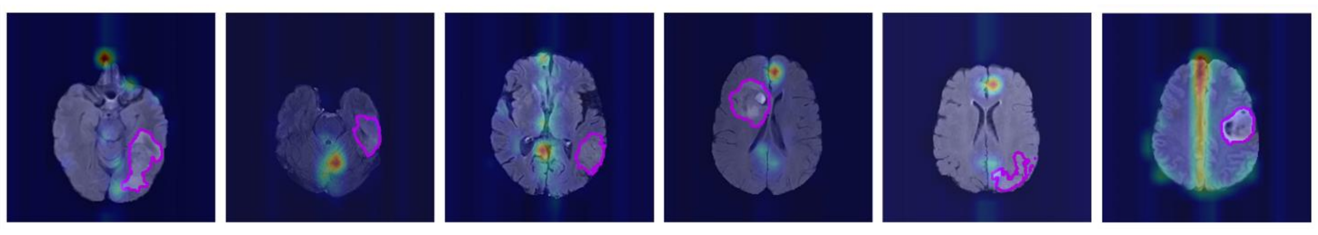
Results

Qualitative Results (BraTS2020 Training Set)

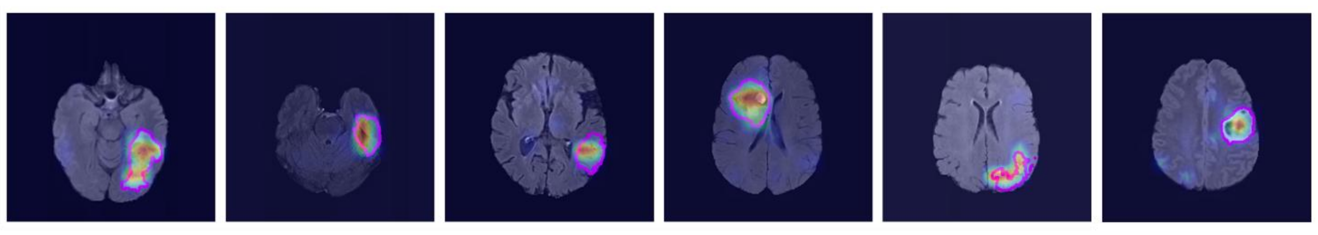
Magenta Contour – Ground-truth lesion annotations

Heatmaps – Explainability maps derived from the model

ViT Baseline: Predictions based on irrelevant features



LGM-ViT: Predictions based on pertinent features



Both models accurately predicted all six examples as **positive** during training.

Quantitative Results

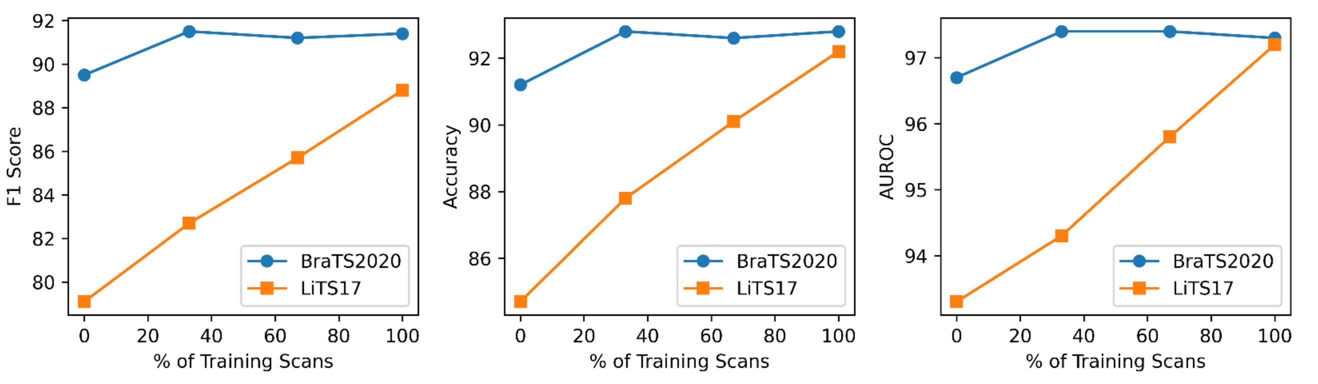
Binary classification evaluation on the ViT-B/16 [4] model:

Dataset	Method	F1 Score	Accuracy	AUROC	AP	Cohen's Kappa
BraTS2020 (lesion)	Baseline	89.5	91.2	96.7	96.5	81.9
	GradMask[6]	89.8	91.4	96.7	96.6	82.4
	RobustViT[3]	89.8	91.3	96.9	96.8	82.2
	RES-G[5]	90.3	91.8	96.9	96.8	83.1
	RES-L[5]	89.6	91.1	96.6	96.6	81.8
	LGM-ViT(Ours)	**91.4**	**92.8**	**97.3**	**97.4**	**85.3**
LiTS17 (liver)	Baseline	79.1	84.7	93.3	90.1	67.0
	GradMask[6]	81.6	87.1	93.7	90.8	71.7
	RobustViT[3]	80.2	86.6	93.3	89.8	70.0
	RES-G[5]	82.0	87.4	94	90.1	72.3
	RES-L[5]	80.3	85.5	92.6	88.1	68.8
	LGM-ViT(Ours)	**88.8**	**92.2**	**97.2**	**96**	**82.8**

Binary classification evaluation for LGM-ViT with different attribution methods on the LiTS17 test set:

Method	F1 Score	Accuracy	AUROC	AP	Cohen's Kappa
None (Baseline)	79.1	84.7	93.3	90.1	67
Rollout Attention[1]	86.2	90.3	96.1	94	78.7
GAE[2]	84.3	88.6	95.4	93.2	75.4
Attention-based Map	82.8	87.9	94.3	92.3	73.5
Embedding-based Map	87.5	91.3	96.3	95.2	80.8
EAFEM	**88.8**	**92.2**	**97.2**	**96**	**82.8**

Performance metrics for LGM-ViT as a function of the percentage of training scans used for localization supervision during training:



Conclusions

- **Challenging medical imaging datasets: Localization supervision works!**
- **Localization supervision on a *small subset* of the data can be enough!**
- **Our approach is not limited to binary classification, and not confined to the medical domain.**

Contact

Sagi Ben Itzhak
Tel Aviv University
Email: sagib2@mail.tau.ac.il

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprintarXiv:2005.00928 (2020)
2. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpret-ing bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVFInternational Conference on Computer Vision. pp. 397–406 (2021)
3. Chefer, H., Schwartz, I., Wolf, L.: Optimizing relevance maps of vision transform-ers improves robustness. Advances in Neural Information Processing Systems 35,33618–33632 (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner,T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image isworth 16x16 words: Transformers for image recognition at scale. arXiv preprintarXiv:2010.11929 (2020)
5. Gao, Y., Sun, T.S., Bai, G., Gu, S., Hong, S.R., Liang, Z.: Res: A robust frame-work for guiding visual explanation. In: Proceedings of the 28th ACM SIGKDDConference on Knowledge Discovery and Data Mining. pp. 432–442 (2022)
6. Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Gradmask: Reduce overfitting byregularizing saliency. arXiv preprint arXiv:1904.07478 (2019)