Random Token Fusion for Multi-View Medical Diagnosis

Jingyu Guo, KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science

Abstract

In multi-view medical diagnosis, deep learning-based models often fuse information from different imaging perspectives to improve diagnostic performance. However, existing approaches are prone to overfitting and rely heavily on view-specific features, which can lead to trivial solutions. In this work, we introduce Random Token Fusion (RTF), a novel technique designed to enhance multi-view medical image analysis using vision transformers. By integrating randomness into the feature fusion process during training, RTF addresses the issue of overfitting and enhances the robustness and accuracy of diagnostic models without incurring any additional cost at inference. We validate our approach on standard mammography and chest X-ray benchmark datasets through extensive experiments.

Framework

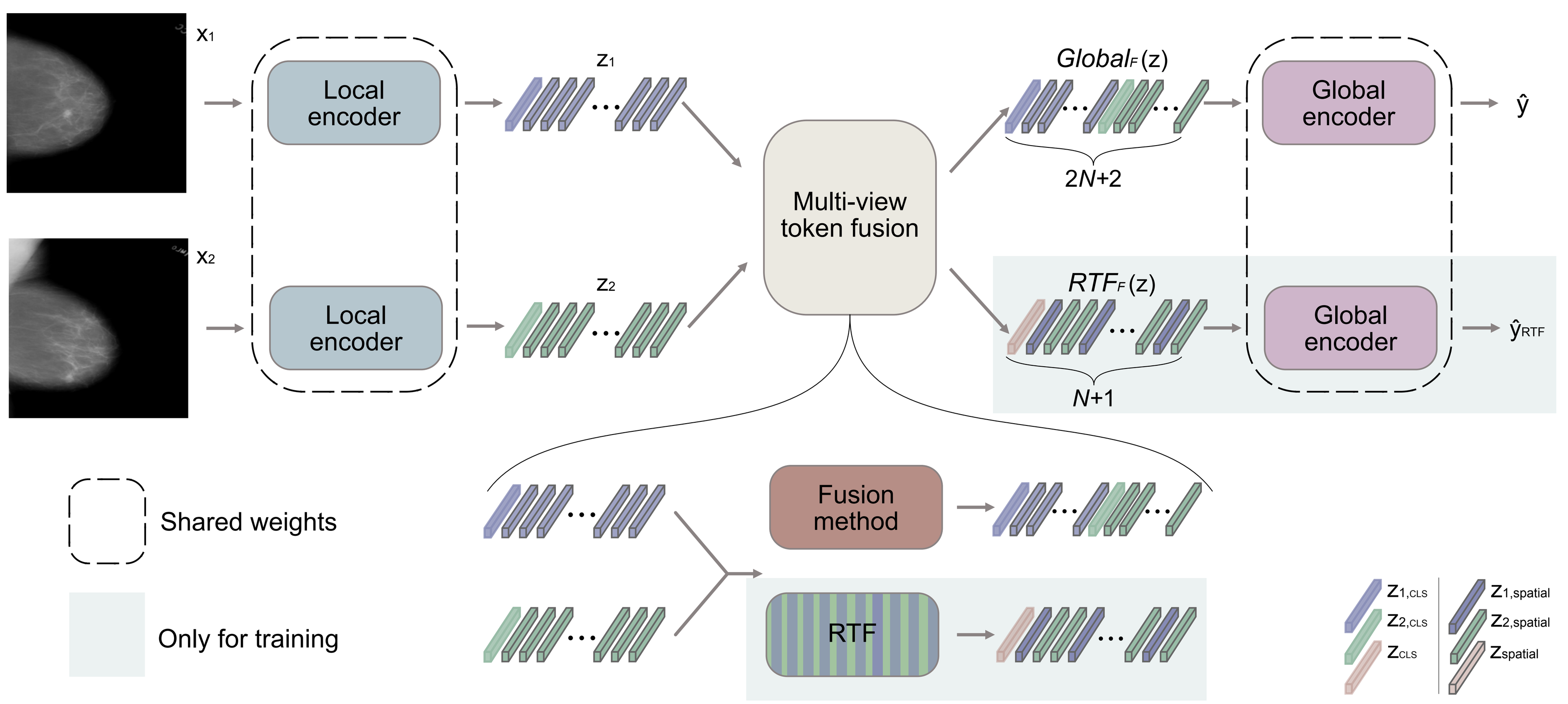

Figure 1: *Multi-view ViTs with Random Token Fusion (RTF). RTF utilizes a local encoder to generate representations of different views, followed by a token fusion module. This module divides the feature fusion into two distinct branches. One branch uses some strategy to merge all tokens from both views, while the other randomly drops spatial tokens from each view before mixing them. The fused tokens are processed by a global encoder, which produces two types of predictions: one for the global tokens and one for the RTF tokens. During training, the loss for both branches is minimized towards the same task. After training, RTF tokens are not generated, they are merged using the model's fusion method and passed to the global encoder for inference.*

Integrating information occlusion and mixing into the training process has been a proven method to combat overfitting and improve robustness [7,8,9]. RTF randomly fuses tokens from different views, introducing variability in the fused representation, which acts as a regularizer. This compels the network to capture dependencies between patches originating from different views, **preventing the model from overfitting to view-specific features**.

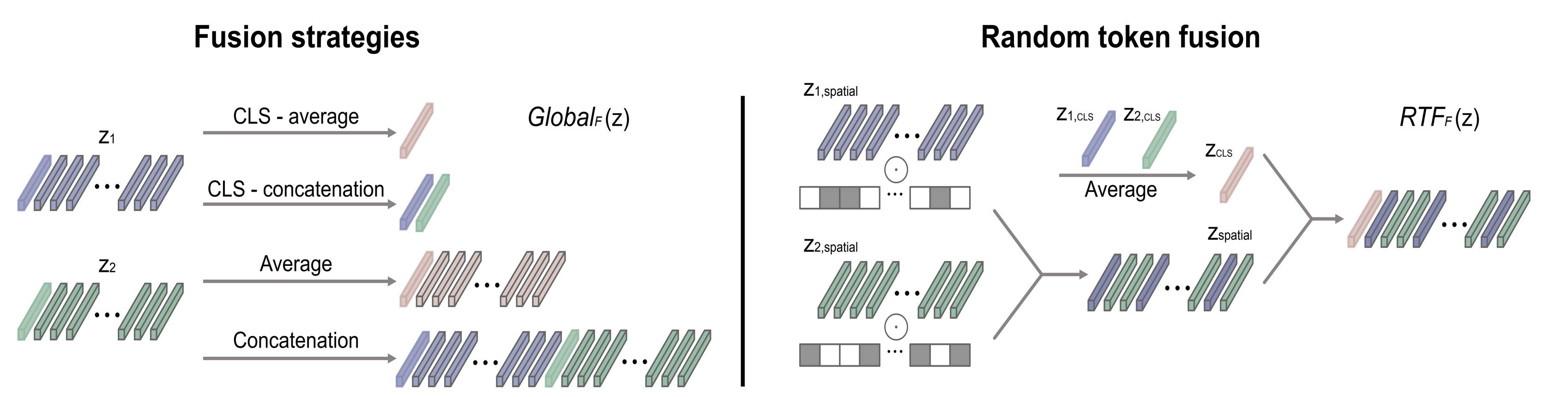

Figure 2: *Illustration of different fusion strategies.* ***(Left)*** *Common fusion strategies to fuse the features (tokens) of different views in ViTs.* ***(Right)*** *The proposed random token fusion (RTF) strategy. In RTF, we randomly drop spatial tokens from both images and combine the remaining ones, augmenting the representations during training.*

RTF can be **seamlessly integrated with existing multi-view fusion strategies** for vision transformers (ViTs), enriching an existing model's feature space without requiring any modification to the inference process. By **incorporating randomness** into the token fusion process, RTF also **encourages the model to learn robust and generalized features** from all views, ensuring that the fused representation captures the most informative features.

Results

RTF enhances multi-view fusion

Training with RTF consistently improves performance across all configurations (Table 1). The extent of improvement varies with the dataset and model size. CBIS-DDSM appears to gain more from RTF, particularly for larger ViT variants. We hypothesize that this is due to the regularization effects of RTF and the smaller size of the dataset, as higher-capacity models are more prone to overfitting.

Table 1: *AUC performance on CBIS-DDSM (**left**) and CheXpert (**right**), showing the effect of using multiple views with and without RTF for different model sizes and fusion strategies.*

Method	RTF Used	ViT Tiny	ViT Small	ViT Base
Average	No	0.798 ± 0.003	0.803 ± 0.008	0.813 ± 0.004
	Yes	$\mathbf{0.802 \pm 0.001}$	$\mathbf{0.809 \pm 0.002}$	$\mathbf{0.825 \pm 0.005}$
CLS_{cat}	No	0.796 ± 0.002	0.802 ± 0.006	0.814 ± 0.007
	Yes	$\mathbf{0.801 \pm 0.001}$	$\mathbf{0.811 \pm 0.008}$	$\mathbf{0.826 \pm 0.004}$
Concat	No	0.798 ± 0.003	0.803 ± 0.003	0.814 ± 0.004
	Yes	$\mathbf{0.802 \pm 0.003}$	$\mathbf{0.815 \pm 0.001}$	$\mathbf{0.830 \pm 0.002}$

Method	RTF Used	ViT Tiny	ViT Small	ViT Base
Average	No	0.798 ± 0.003	0.803 ± 0.008	0.813 ± 0.004
	Yes	$\mathbf{0.802 \pm 0.001}$	$\mathbf{0.809 \pm 0.002}$	$\mathbf{0.825 \pm 0.005}$
CLS_{cat}	No	0.796 ± 0.002	0.802 ± 0.006	0.814 ± 0.007
	Yes	$\mathbf{0.801 \pm 0.001}$	$\mathbf{0.811 \pm 0.008}$	$\mathbf{0.826 \pm 0.004}$
Concat	No	0.798 ± 0.003	0.803 ± 0.003	0.814 ± 0.004
	Yes	$\mathbf{0.802 \pm 0.003}$	$\mathbf{0.815 \pm 0.001}$	$\mathbf{0.830 \pm 0.002}$

RTF outperforms SOTA fusion methods

RTF outperforms other fusion methods on both datasets, showing its efficacy in multi-view medical diagnosis. Notably, RTF can be used in conjunction with transformer-based methods, such as ViT-Average [11] and MVT [12,13], for further enhanced performance.

Table 2: *Comparison vs. SOTA methods on CBIS-DDSM (**left**) and CheXpert (**right**).*

Method	CBIS-DDSM
ResNet50	0.724 ± 0.007
Shared ResNet	0.735 ± 0.014
PHResNet50	0.739 ± 0.004
MVT	0.803 ± 0.003
CVT	0.803 ± 0.007
ViT-Average	0.803 ± 0.008
RTF	$\mathbf{0.815 \pm 0.001}$

Method	CheXpert
MVC-NET	0.813 ± 0.005
MVCNN	0.815 ± 0.004
CVT	0.834 ± 0.002
MVT	0.843 ± 0.004
ViT-Average	0.844 ± 0.004
MV-HFMD	0.845 ± 0.002
RTF	$\mathbf{0.849 \pm 0.001}$

References

[1] Wu, N., et al. "Improving the ability of deep neural networks to use information from multiple views in breast cancer screening." Medical Imaging with Deep Learning. PMLR, 2020.
[2] Wang, W., et al. "What makes training multi-modal classification networks hard?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[3] Feigin, D. S., et al. "Lateral chest radiograph: a systematic approach." Academic radiology 17.12 (2010): 1560-1566.
[4] Hashir, M., et al. "Quantifying the value of lateral views in deep learning for chest x-rays." Medical Imaging with Deep Learning. PMLR, 2020.
[5] Raoof S., et al. "Interpretation of plain chest roentgenogram." Chest 141.2 (2012): 545-558.
[6] Irvin, J., et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. 2019.
[7] DeVries, T. "Improved Regularization of Convolutional Neural Networks with Cutout." arXiv preprint arXiv:1708.04552 (2017).
[8] Zhang, H., et al. "mixup: Beyond Empirical Risk Minimization." International Conference on Learning Representations. 2018.
[9] Liu, Y., et al. "Patchdropout: Economizing vision transformers using patch dropout." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[10] Sawyer-Lee, R., et al. "Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM)" on The Cancer Imaging Archive. (2016).
[11] Nguyen, H. TX., et al. "A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms." 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022.
[12] Chen, S., et al. "Mvt: Multi-view vision transformer for 3d object recognition." arXiv preprint arXiv:2110.13083 (2021).
[13] Chen, X., et al. "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms." Diagnostics 12.7 (2022): 1549.

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM