Learning Contrastive Multimodal Fusion with Improved Modality Dropout for Disease Detection and Prediction

Yi Gu[1,2] Kuniaki Saito[1] Jiaxin Ma[1]✉

1 OMRON SINIC X Corporation 2 Nara Institute of Science and Technology

OMRON SINIC X

NARA INSTITUTE of SCIENCE and TECHNOLOGY

Introduction

Medical diagnoses increasingly leverage **multimodal data.**

Data and modality **missingness** happens in training and inference, practically.

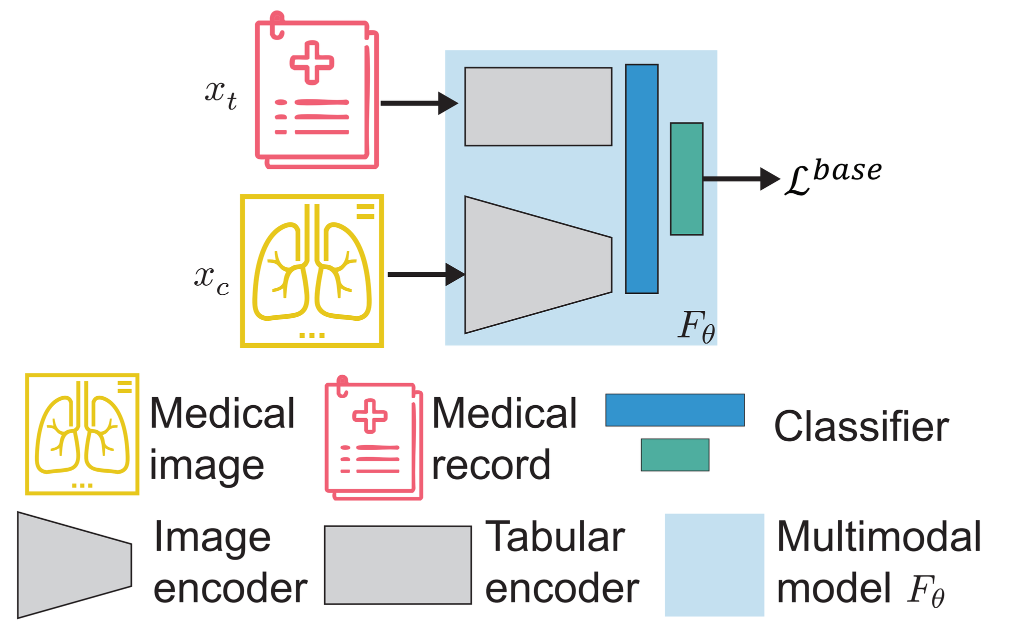
Multimodal learning is often **costly.**

Effective multimodal learning algorithms require deeper exploration.

Challenge remain.

Motivation

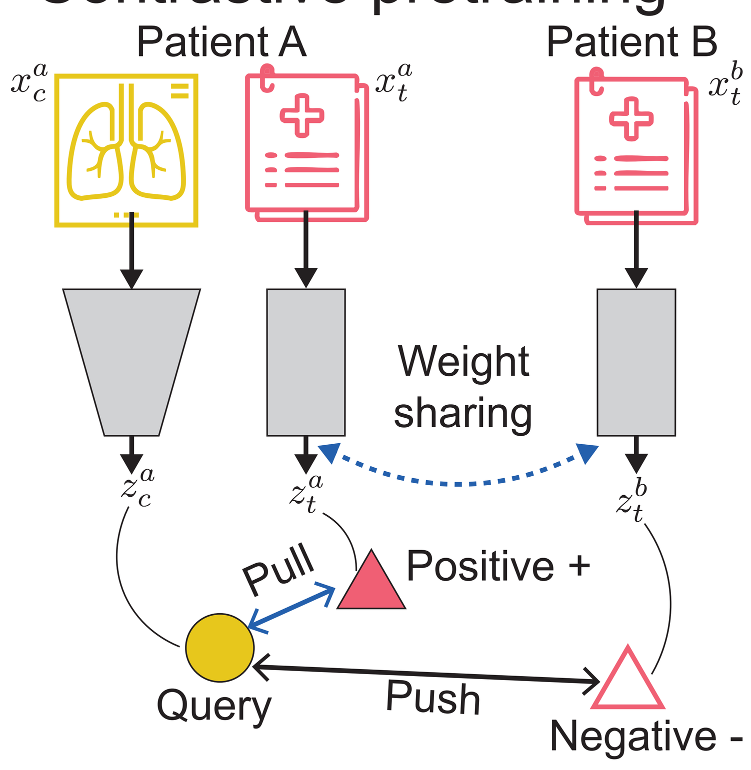
▪Neural fusion multimodal training



$p(y|x_c, x_t) = F_\theta(x_c, x_t)$ $\mathcal{L}^{base} = -\log p(y|x_c, x_t)$

Could not handle data missingness.

▪Contrastive pretraining



z_c : Image representation

z_t : Tabular representation

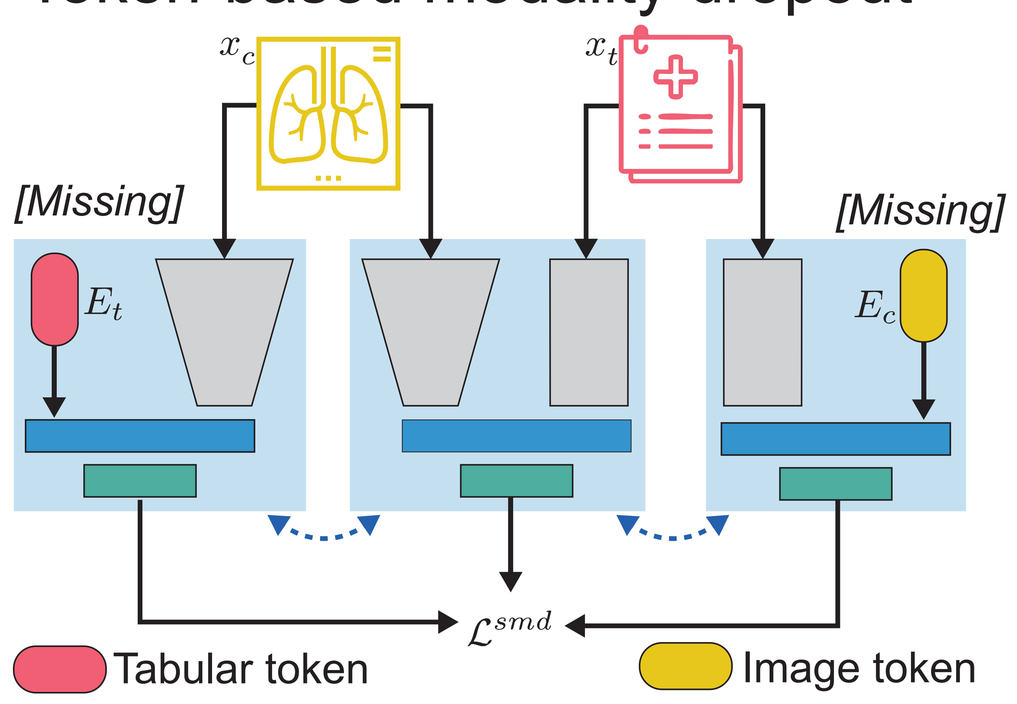
$$\mathcal{L}^{con} = -\log \frac{\exp(z_c^a \cdot z_t^a/\tau)}{\sum_i^{\{a,b\}} \exp(z_c^a \cdot z_t^i/\tau)}$$

τ: Learnable tempreature

Only learns unimodal representation.

Method

▪Token-based modality dropout



E_t : Tabular data missingness token

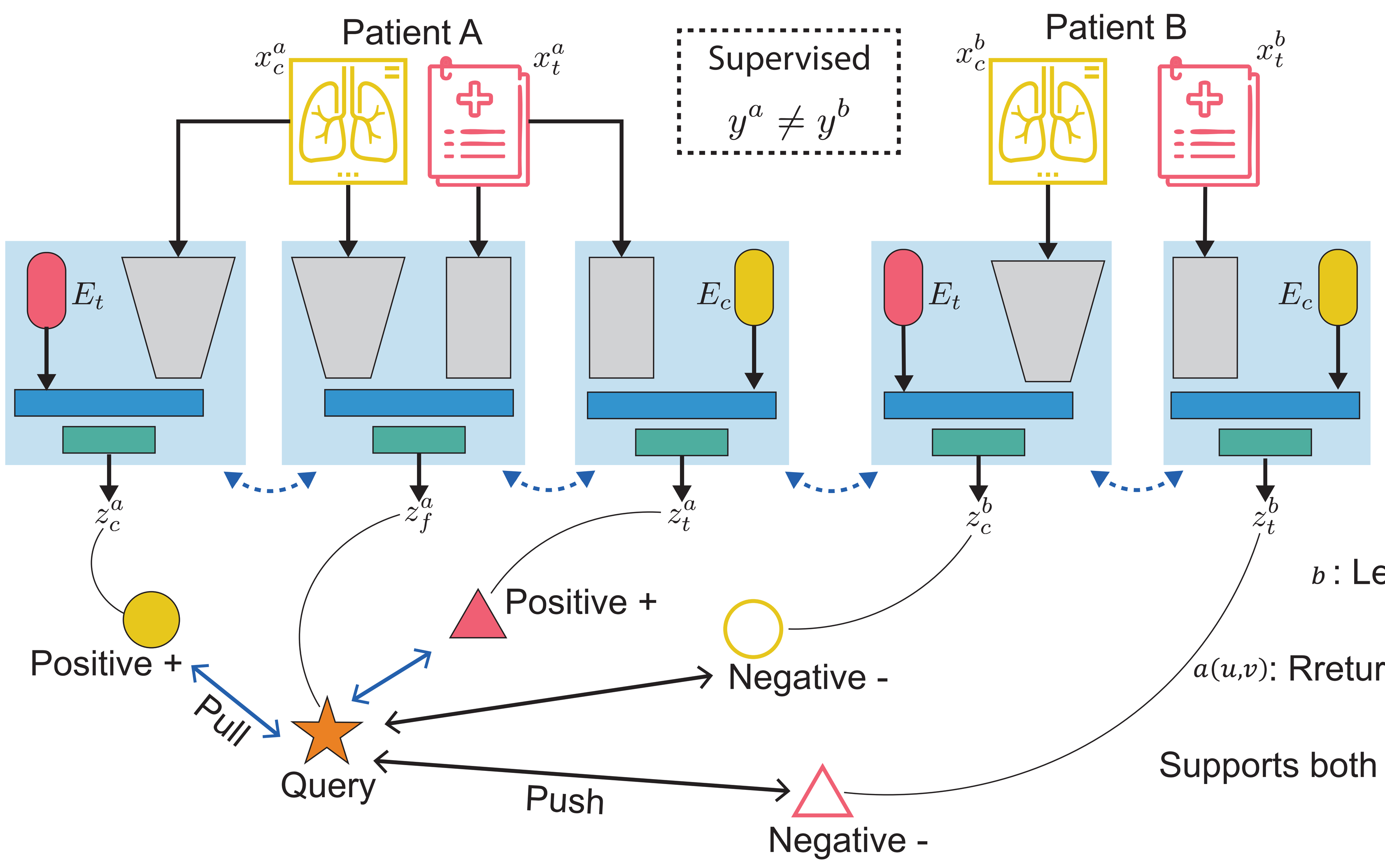
E_c : Image missingness token

$p(y|x_c) = F_\theta(x_c, E_t)$ $p(y|x_t) = F_\theta(E_c, x_t)$

$$\mathcal{L}^{smd} = -\log p(y|x_c, x_t) - \sum_{i \in \{c,t\}} p(y|x_i)$$

Modality dropout with missinngness token.

▪Contrastive multimodal fusion pretraining



z_f : Image-tabular fused representation

$$\mathcal{L}_i^{con} = -\sum_{u \in N} \sum_{v \in N} \log \frac{1}{1 + e^{a(u,v)\left(-t z_f^u \cdot z_i^v + b\right)}}$$

$$\hat{\mathcal{L}}^{con} = \mathcal{L}_c^{con} + \mathcal{L}_t^{con}$$

N : Set of patients

t : Learnable rescale parameter

b : Learnable shift parameter

$a(u,v)$: Rreturns 1 for positive pair and -1 for negative pair

Supports both supervised and unsupervised training

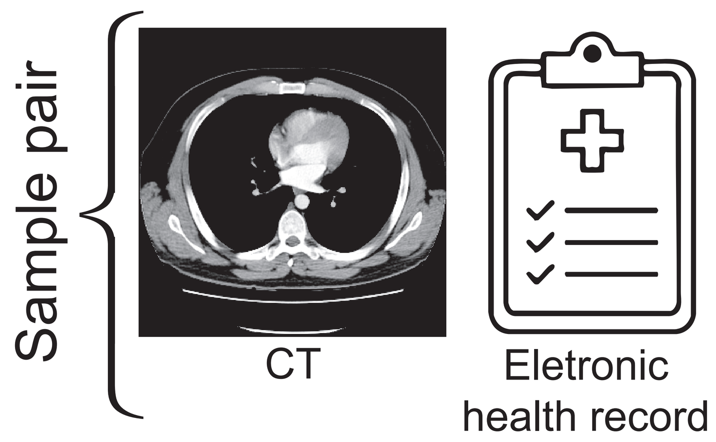
Leveraging fused mutimodal representation in contrastive learning.

Results

▪Experiment datasets: Public chest CT datasets with EHR

PE-dataset with **1,837** CT images with **pulmonary embolism** labels.

NLST dataset with **64,117** CT images with **future cancer** labels.



▪PE-dataset result table

	Inference CT	Inference Table	AUROC↑	AP↑	AURC↓	MCC↑	F-score↑
PENet[1]	✓	-	0.758	0.609	0.475	0.207	0.538
PENet †[1]	✓	-	0.778	0.680	0.442	0.379	0.614
Ours †	✓	-	**0.801**	**0.724**	**0.422**	**0.451**	**0.647**
ElasticNet[2]	-	✓	**0.758**	**0.581**	**0.487**	**0.370**	0.564
FT-Transformer[3]	-	✓	0.745	0.539	0.510	0.351	**0.597**
Ours †	-	✓	0.751	0.558	0.499	0.352	0.594
DAFT[4]	✓	✓	0.739	0.536	0.511	0.334	0.595
DAFT †[4]	✓	✓	0.629	0.459	0.566	0.174	0.437
DAFT-64[4]	✓	✓	0.700	0.459	0.566	0.259	0.561
DAFT-64 †[4]	✓	✓	0.616	0.479	0.560	0.139	0.434
TabAttention[5]	✓	✓	0.738	0.551	0.505	0.271	0.565
RadFusion[1]	✓	✓	0.811	0.676	0.438	0.294	0.572
RadFusion †[1]	✓	✓	0.819	0.716	0.422	0.418	0.633
RadFusion (FT)	✓	✓	0.803	0.642	0.454	0.288	0.567
RadFusion (FT) †	✓	✓	0.815	0.707	0.426	0.425	0.640
Ours †	✓	✓	**0.842**	**0.775**	**0.397**	**0.499**	**0.676**

▪NLST-dataset result table

	Cancer in 2 years AUROC↑	Cancer in 2 years AP↑	Cancer in 2 years AURC↓	Cancer in 1 year AUROC↑	Cancer in 1 year AP↑	Cancer in 1 year AURC↓
CT foundation model [6]	0.729	0.070	0.956	0.700	0.050	0.972
Ours (CT only)	0.732	0.068	0.956	0.780	0.068	0.969
ElasticNet	0.808	0.199	0.938	0.830	0.132	0.962
FT-Transformer	0.837	0.246	0.933	0.925	**0.279**	**0.949**
Ours	**0.857**	**0.247**	**0.931**	**0.926**	**0.279**	**0.949**

▪Ablation study result table (AUC)

Training loss	Modality token	Pretraining loss	Image-tabular inference PE	Image-tabular inference $NLST_2$	Image-tabular inference $NLST_1$	Image-only inference PE	Image-only inference $NLST_2$	Image-only inference $NLST_1$
$\mathcal{L}^{base}$	/	-	0.837	0.847	0.919	/	/	/
$\mathcal{L}^{md}$	-	-	0.836	0.850	0.917	0.797	0.714	0.707
$\mathcal{L}^{md}$	✓	-	0.840	0.855	0.915	0.796	0.728	0.741
$\mathcal{L}^{smd}$	-	-	0.838	0.851	**0.926**	0.800	0.716	0.738
$\mathcal{L}^{smd}$	✓	-	0.840	0.855	0.920	0.800	0.722	0.751
$\mathcal{L}^{smd}$	✓	$\mathcal{L}^{con}$	**0.842**	0.856	**0.926**	0.800	0.726	0.771
$\mathcal{L}^{smd}$	✓	$\hat{\mathcal{L}}^{con}$	**0.842**	**0.857**	**0.926**	**0.801**	**0.732**	**0.780**

Ours achieved the best scores under the image-only and image-tabluar inference setting.

Improvements from **individual components** were verified through an ablation study.

Conclusion

Proposed missingness token in modality dropout for a better data **missingness handling.**

leveraged fused representation in contrastive learning for a better **multimodal binding.**

reference

[1] Zhou, Y., et al. (2021): RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR. *arXiv*.

[2] Zou, H & Hastie, T. (2005): Regularization and Variable Selection Via Elastic Net. *J. R. Stat. Soc. Ser. B Methodol.*

[3] Gorishniy, Y. et al. (2021): Revisiting deep learning models for tabular data. *NuerIPS*.

[4] Polsterl, S. et al. (2021): Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. *MICCAI*.

[5] Grzeszczyk, M.K. (2023), et al.: Tab Attention: Learning Attention Conditionally on Tabular Data. *MICCAI*.

[6] Yang, L., et al. (2024): Advancing multimodal medical capabilities of gemini. *arXiv*.

Code Yi Gu's website

MICCAI2025

gu.yi.gu4@naist.ac.jp jiaxin.ma@sinicx.com