Advanced Architectures

Tanveer Syeda-Mahmood

Stanford MEDICINE

Today's topics

- Fact-checking models
- Multimodal graph neural networks
- State-space models
- Memory-based models

Responsible radiology reporting

- Accurately reporting what is found in image (presence)
  - Core finding itself
  - Its location
  - Severity
- Mention some findings not found in image (absence)
  - Which ones to report?
- Avoid irrelevant other descriptions
- Avoid typos

Ground Truth Reports

overall impression : Left hilar opacity may represent primary lung mass. Left hilar opacity. Left port.

Small left effusion. Pleuroparenchymal opacities at the left lung base. Wires external to patient. Surgical clips superior to the left clavicle.

Automatically Generated Reports

lines and tubes: there is a right chest port catheter terminating in the lower lungs: there is a large mass in the right lower lobe overall: left lower lobe opacities concerning for pneumonia

lines and tubes:
r picc tip at the svc wires external to patient
lungs:
right lower lobe consolidation
overall:

Errors in automated radiology reports

Present findings: enlarged cardiac silhouette, pacer, calcifications, sternotomy wires, mediastinal clips, clear lungs
Absent findings: pleural effusion, vascular congestion

Ground truth report

MEDICAL CONDITION:
78 year old woman with chest pain, hx of chf
REASON FOR THIS EXAMINATION:
please evaluate for edema
FINAL REPORT
CHEST, TWO VIEWS:
HISTORY: 78-year-old female with chest pain. History of congestive failur
Question pulmonary edema.

FINDINGS: PA and lateral views of the chest are compared to previous exam from .
Compared with prior, there has been no significant interval change. The lungs remain clear. There is no pleural effusion. There is no pulmonary vascular engorgement. Cardiac silhouette is enlarged, but stable in configuration. Biventricular pacing device again seen with multiple leads i stable positions. Atherosclerotic calcifications seen throughout the aorta Median sternotomy wires and mediastinal clips again noted.

IMPRESSION: No acute cardiopulmonary process.

Match to GT Mismatches

X-ray GPT:

In this patient who has an upper lobe nodule and associated pain complaints, there is no indication of pleural effusion and cardiac failure. However, edema can be present, which should be evaluated to exclude a pulmonary disease. The patient has mild pulmonary fibrosis and may have underlying COPD which should also be considered. The findings suggest cardiac failure and the patient requires diuretics for treatment. The ECG appears normal and the dual lead AID pacing device remains intact.

RGRG:

The lungs are clear. No acute cardiopulmonary process. The cardiomediastinal silhouette and hilar contours are normal. There is no focal consolidation, effusion, or pneumothorax. There are no acute osseous abnormalities. Moderate cardiomegaly is unchanged. Left-sided pacemaker leads terminate in the right atrium and right ventricle.

GPT4o:
FINDINGS:
Lines/Tubes: Pacemaker leads are in place.
Lungs: Mild interstitial markings observed; no acute infiltrates.
Pleura: No significant pleural effusion detected.
Heart and Mediastinum: Enlarged cardiac silhouette consistent with known CHF.
Bones and soft tissues: No acute fractures or bone abnormalities detected.

Detecting errors in RRG reports

- Agnostic to RRG models
  - Why?
- Should work at inference time without re-training RRG models
- Types of errors to be modeled
- Role of synthetic data generation
- Building phrase-grounded discriminative models

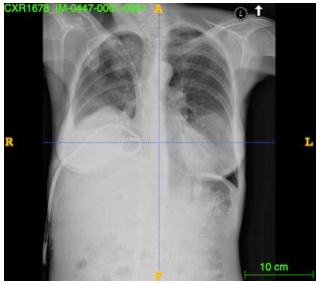
Phrase-grounded Fact-checking for Automatically Generated Chest X-ray Reports (MICCAI 2025)

Error modeling in automated reports

Types of errors found in reports

- Correct Finding (No error)
- Incorrect Finding
  - Irrelevant finding
  - Reverse finding
  - Exchange finding
- Incorrect location
- Missing or other Findings

Chest X-ray



Associated GT report

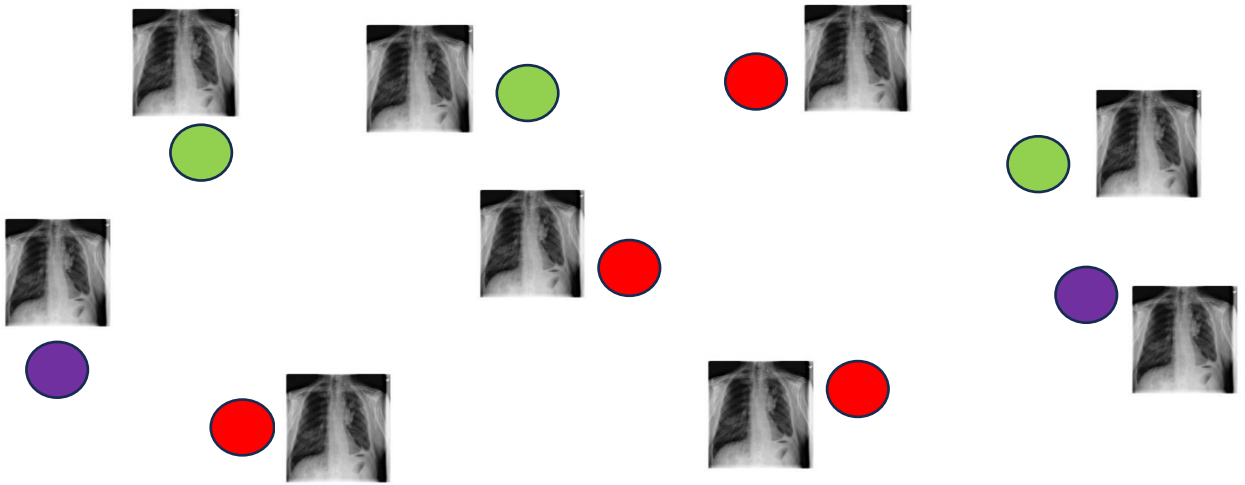
The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Synthetic data generation – key idea

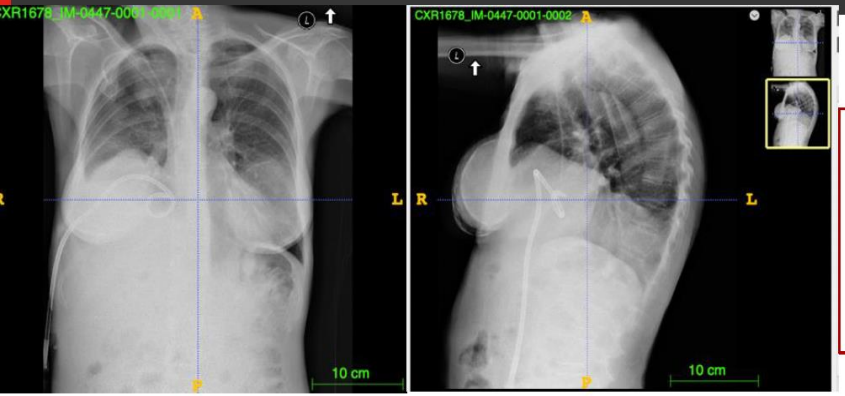
- Collect labeled image-finding pairs simulating errors

Synthesize Image-Text Data

- Representative of typical errors in automated reports
- Realistic and possible combinations verifiable through ground truth
- Robust to written variations



Synthetic data generation to simulate errors



Original Report

The heart is normal in size. Right chest tube tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Missed finding

The heart is normal in size.
There is no pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion.
Vague opacities are noted in the right upper lobe, xxxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.

Finding missed: “chest tube position”

Added finding

The heart is normal in size.
Right chest xxxx tip is again seen at the cavoatrial junction.
There is no pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion.
Vague opacities are noted in the right upper lobe, xxxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.
There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lung xxxx..

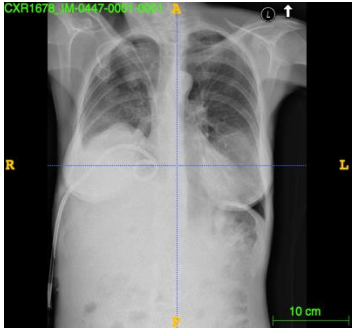
Reversed finding

The heart is normal in size.
Right chest xxxx tip is again seen at the cavoatrial junction.
There is pneumothorax.
There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxx from prior study.
These may be related to overlying rib lesions versus true pulmonary nodules.
The left lung appears grossly clear.
Drainage catheter seen overlying the right upper quadrant.

Requires extracting findings from reports

Synthesizing image-text pairs

Synthetic Data Generation



- *Many many combinations possible!*
- *Not all of them are realistic.*
- *How to filter?*
  - *Based on findings provided*
  - *Avoid repetitions*
  - *Remove Contradictions*

R1

The heart is normal in size. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion.

- Normal heart
- No pneumothorax
- Elevated hemidiaphragm
- Pleural effusion

R2

The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

- Clear lung
- Chest tube

R3

There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lung. There is a small left pleural effusion/pneumothorax.

- ILD
- Pulmonary fibrosis
- Pleural effusion
- Pneumothorax

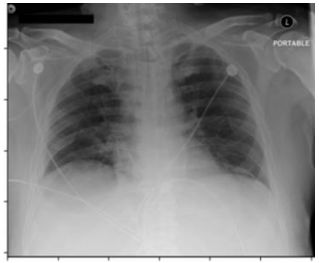
Real

Fake

Drop pleural effusion
Drop pneumothorax

Synthesizing Data – Labeling findings

Synthetic Data Generation



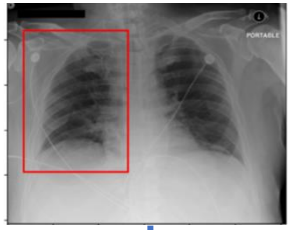
Mild basilar {atelectatic changes} without evidence of acute pneumonia or vascular congestion. There may well be a small right pleural effusion.

There may well be a small right pleural effusion

FFL: Anatomical finding|yes|pleural effusion|lung|right

Short FFL Yes|Pleural effusion

Left lung



Anatomical location <177,395,1146,1569>

Normalized location <.06,0.15,0.37,0.62>

INPUT

Yes|Pleural effusion

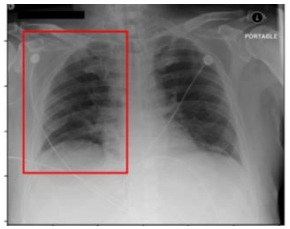
LABEL

<.06,0.15,0.37,0.62,1>

1=Real,
0=Fake

Synthesizing Data – Perturbing presence findings

Synthetic Data Generation

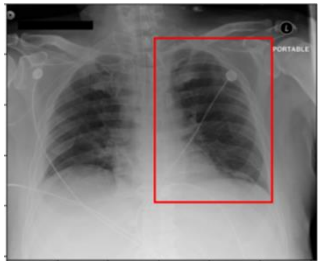


Yes|Pleural effusion

<.06,0.15,0.37,0.62,1>

There may well be a small right pleural effusion.

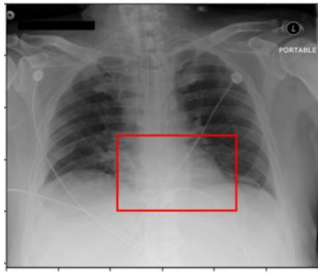
Finding is correct but location is incorrect



Yes|Pleural effusion

<0.47, 0.13, 0.38,0.64,0>

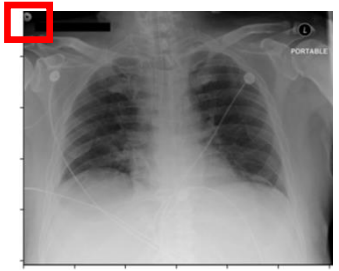
Exchange Finding



Yes|Cardiomegaly

<0.35, 0.5,0.37,0.28,0>

Reverse Finding

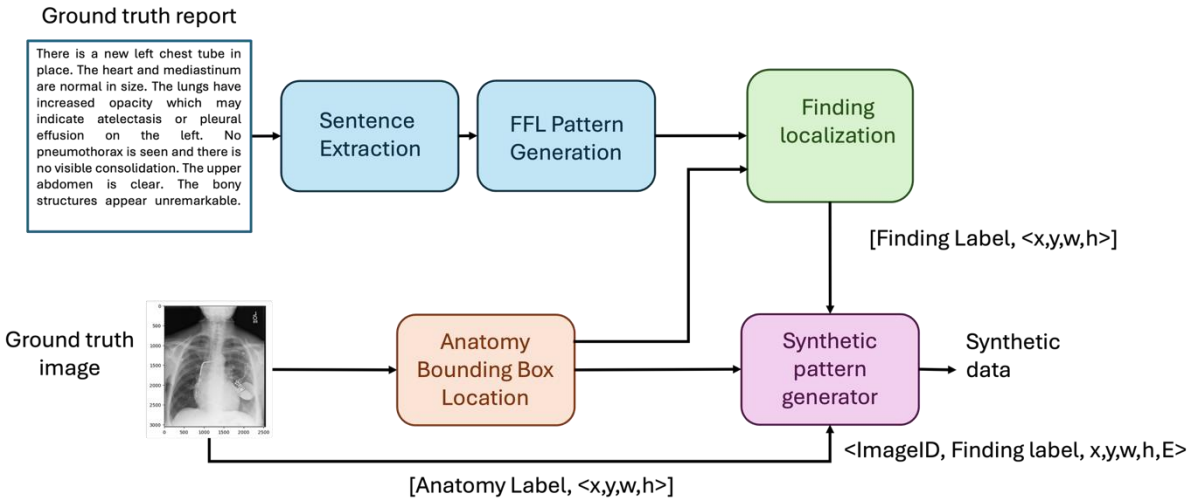


No| Pleural effusion

<0,0,0,0,0>

Synthetic data generation

Synthetic Data Generation



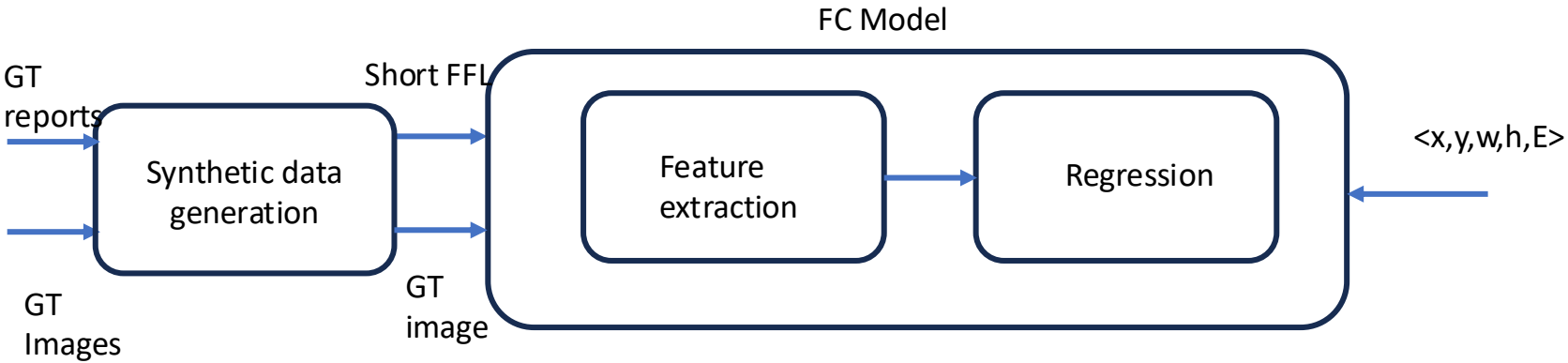
Millions of samples can be generated!

Image	Short FFL	Location	Finding/location correct?
	Yes|pleural effusion	<0.06,0.15,0.37,0.62>	1
	No|vascular congestion	<0,0,0,0>	0
	Yes|cardiomegaly	<0.35, 0.5,0.37,0.28>	0

Fact-checking model

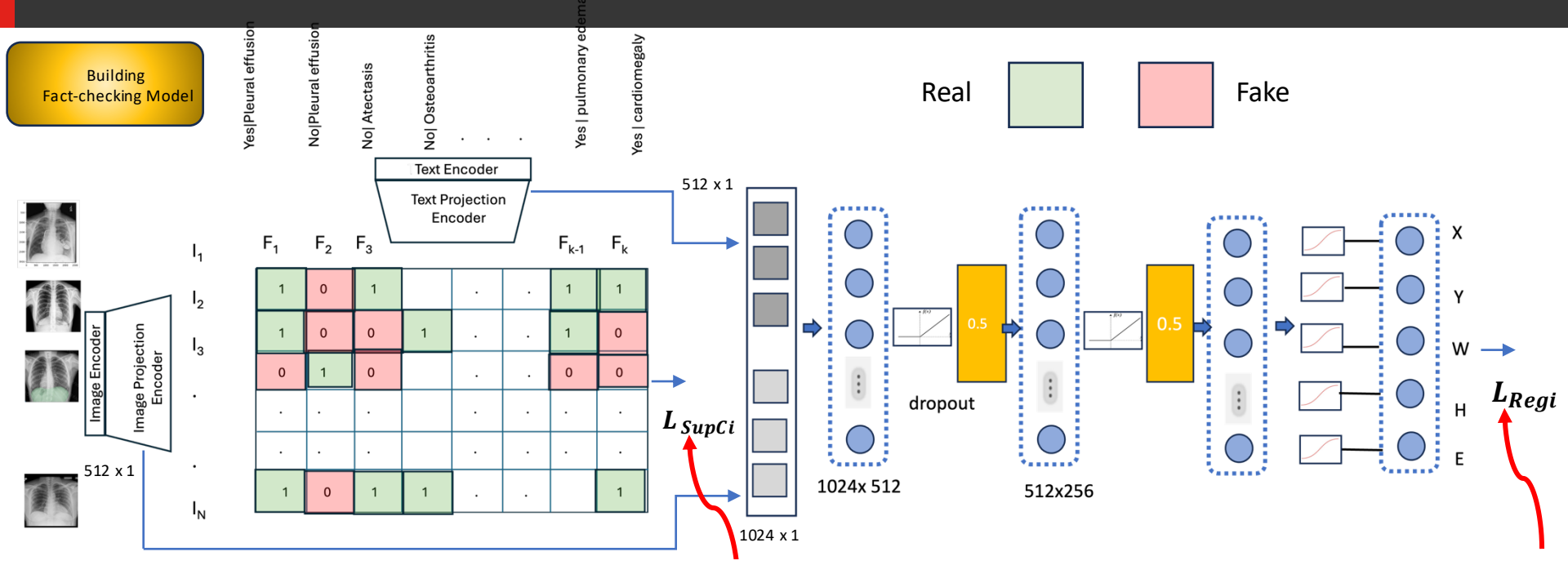
- Discriminative model to separate real from fake pairings of:
  - <images, findings, locations>
- Can a VLM be used to model image and text?
  - How to handle location?

FC Model Design



- How to choose feature extractor?
- How to design a regressor?
- Can this be done as an end-to-end neural network?

FCModel: Supervised Cross-modal Contrastive Regression Network



$$\mathcal{L}_{SupC} = \sum_{i \in I} \frac{-1}{|F_{iReal}|} \sum_{f_{ij} \in F_{iReal}} log \frac{e^{s_i f_{ij}/\tau}}{\sum_{a_{ik} \in F_{iFake}} e^{s_{ia_{ik}}/\tau}}$$

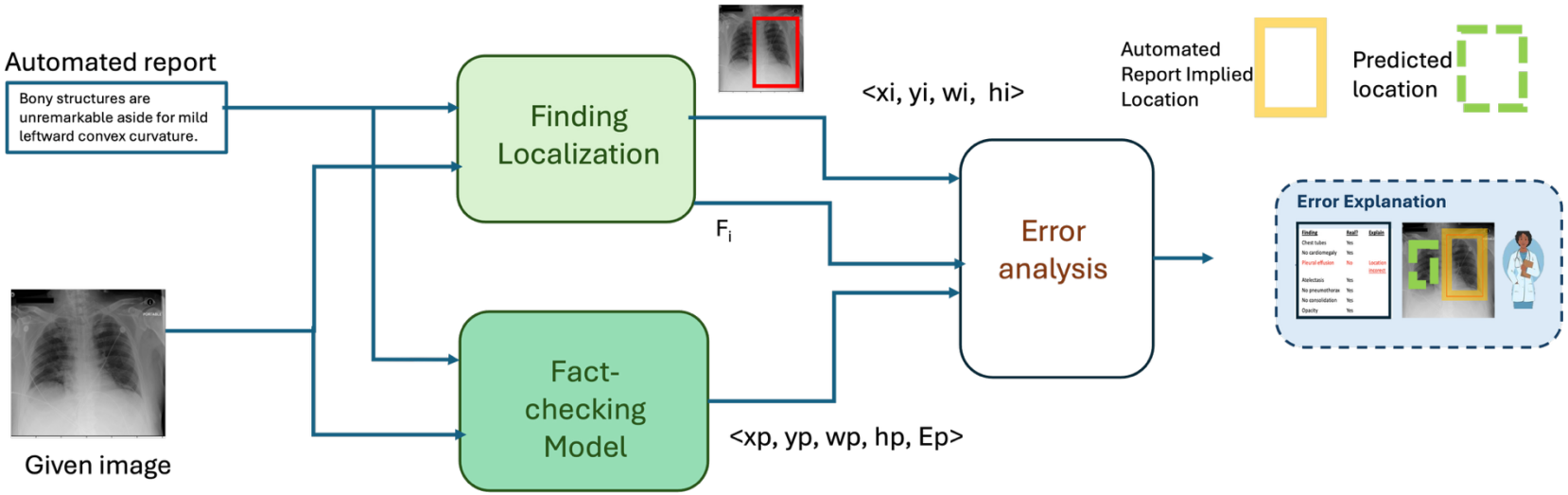
$$\mathcal{L}_{Reg_i} = \underbrace{|Y_{1p} - Y_{1g}|}_{\mathcal{L}_1(Y_{1p}, Y_{1g})} + \underbrace{\frac{|Y_{1p} \cap Y_{1g}|}{|Y_{1p} \cup Y_{1g}|} - \frac{|C_{Y_{1p},Y_{1g}} \backslash Y_{1p} \cap Y_{1g}|}{|C_{Y_{1p},Y_{1g}}|}}_{\mathcal{L}_{\text{giou}}(Y_{1p}, Y_{1g})}$$

$$+ \underbrace{|Y_{1p} - Y_{1g}|^2}_{\mathcal{L}_{\text{mse}}(Y_{1p}, Y_{1g})} - \underbrace{[Y_{2g}\mathbf{log}(Y_{2p}) + (1 - Y_{2g})\mathbf{log}(1 - Y_{2p})]}_{\mathcal{L}_{BCE}(Y_{2p}, Y_{2g})}$$

- Built as an end-to-end trainable network
  - image and text projection encoders
  - concatenated image-text representation
  - Regression on contrasted features
- Findings can be presence or absence findings

Using FCModel in a clinical workflow

Building
Fact-checking Model



Fact-checking examples

Building
Fact-checking Model

Case: Finding and location are correct

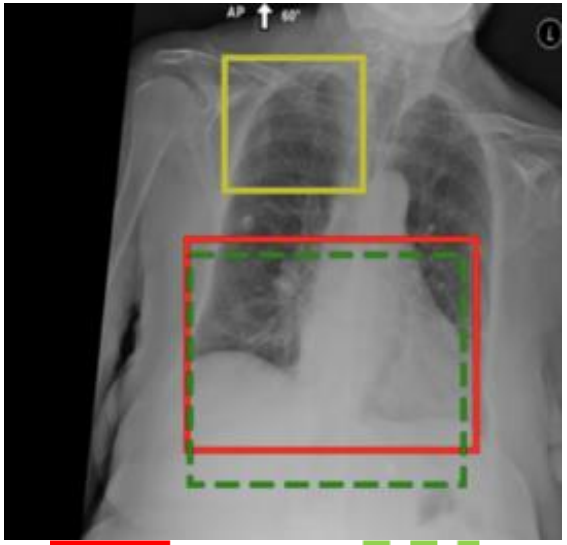
Ground truth sentence	... pronounced alveolar opacities in a perihilar distribution, right greater than left, likely representing pulmonary edema
Automated report sentence	The pulmonary vasculature is not engorged, and the patient has moderate pulmonary edema on the right.
Target finding	Yes|Pulmonary edema
Implied merged anatomical locations	['right hilar structures', 'right lung']
GT location	[0.120, 0.125, 0.366 0.586]
Predicted location	[0.151, 0.100, 0.351, 0.538, 0.428]
Predicted label	1

GT Predicted Indicated

Fact-checking using FCModel

Building
Fact-checking Model

Case: Finding is correct but location is wrong

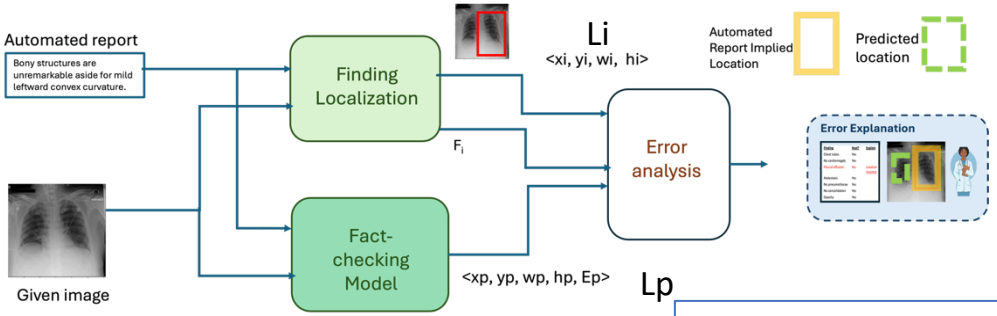


Ground truth sentence	Opacities are present at the lung bases bilaterally.
Automated report sentence	Subtle opacity in the right upper lung.
Target finding	Yes|Opacity
Implied merged anatomical locations	[right upper lung]
GT location	[0.328, 0.419, 0.498, 0.363]
Predicted location	[0.332, 0.442, 0.470, 0.396]
Predicted label	0

GT Predicted Indicated

Error correction using FCModel

Report Correction

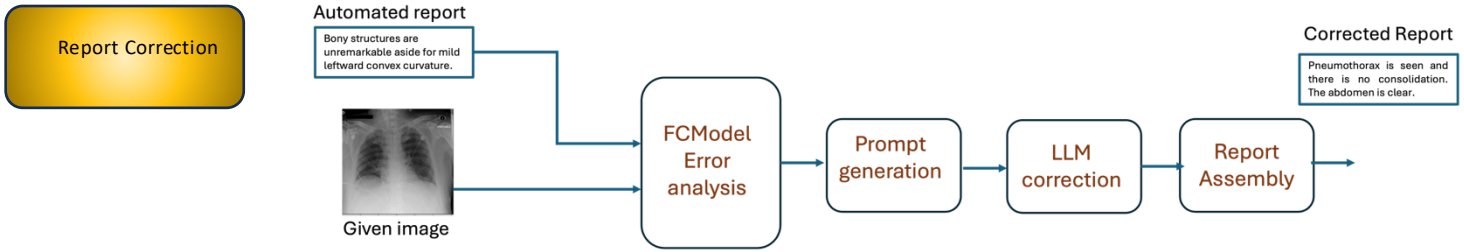


6 combinations of errors -> 3 corrective actions

Fi	Ep	1-IOU(Lp,Li)	Interpretation	Corrective action
Absent	0	>Γ	Finding is either correct or incorrect but location is wrong.	Remove the finding altogether.
Absent	1	<Γ	Finding and its location are correct	No action
Absent	0	<Γ	Finding is incorrect	Flip the finding only. Keep the location.
Present	0	>Γ	Finding is correct or incorrect. Location is wrong	Remove the finding altogether.
Present	1	<Γ	Finding and location are correct	No action
Present	0	<Γ	Finding is incorrect. Location is correct	Flip the finding only. Keep the location.

All other cases : No action

Report correction



Corrective Action	Prompt template
No action	None
Flip the finding. Keep the location	Remove ”no <Fcp >”and add ”yes <Fcp >”in the sentence: < Sp >
Remove the finding altogether.	Remove ” < Fi >” and “< location name (from full FFL corresponding to Fp)>” in the sentence:<Sp>

Report correction examples

Original sentence	Target finding	Corrective action	Prompt	Corrected sentence
Left-sided pleural effusion found and the right atelectasis still remains.	yes|pleural effusion	Remove the finding altogether	Remove ”pleural effusion” from sentence.	Right atelectasis still remains.
The chest x ray image shows no focal consolidation, pulmonary edema, pleural effusion or pneumothorax .	no|pneumothorax	Remove the finding altogether	Remove ”no pneumothorax” from the sentence:	The chest X-ray image shows no focal consolidation, pulmonary edema, pleural effusion.
The chest x ray image shows no focal consolidation, pulmonary edema, pleural effusion or pneumothorax .	no|pneumothorax	Flip finding.	Remove ”no pneumothorax” and add ”yes pneumothorax” in the sentence:	The chest X-ray image shows pneumothorax, but no focal consolidation, pulmonary edema, or pleural effusion.
There is left pleural effusion and pneumothorax	yes|pleural effusion	Remove finding and location	Remove ”left pleural effusion” but keep the ”left pneumothorax” in the sentence:	There is left pneumothorax.
The chest x ray image shows no left focal consolidation.	no|consolidation	Remove the location. Keep the finding.	Remove ”left” from the sentence:	The chest x ray image shows no focal consolidation.

Experimental results - Datasets

Dataset	Patients Train/Val/Test	Images	Findings	Regions	Real/Synth.
CImagenomeS[79]	44,133/6274/12,538	243,311	49	922,295	1.616M/27.047M
CImaGenomeG[79]	288/33/69	461	35	5,477	4,063/23,463
MS-CXR[24]	478/54/114	925	8	2,254	2,247/24,338
ChestXray8[71]	457/51/109	880	8	1,571	1,571/10,137
VinDr-CXR[43]	9,450/1,050/2,250	15,000	23	69,052	47,973/132,632

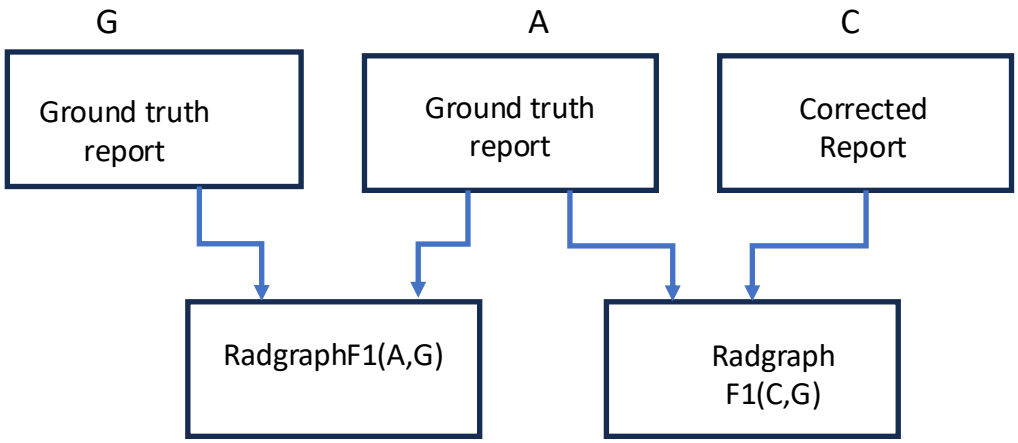
- All datasets clinician validated
- ChestImagenomeS (Silver) used for training the FCModel only
- ChestImagenomeG (Gold) used for other evaluations (comparisons, report correction, etc.)

Findings (Presence and Absence)

1.	Aortic enlargement	10.	Atelectasis	18.	Calcification
2.	Cardiomegaly	11.	Clavicle fracture	19.	Consolidation
3.	Edema	12.	Emphysema	20.	Enlarged pa
4.	ILD	13.	Infiltration	21.	Infiltrate
5.	Lung cavity	14.	Lung cyst	22.	Lung opacity
6.	Mediastinal shift	15.	Nodule/mass	23.	Other lesion
7.	Pleural effusion	16.	Pleural thickening	24.	Pneumothorax
8.	Pulmonary fibrosis	17.	Rib fracture	25.	Pneumonia
9.	No finding				

Largest set of findings considered!

Report correction performance using clinical accuracy



- 7 Report generators evaluated:
  - RGRG
  - XrayGPT
  - GPT4o
  - R2GenGPz
  - CV2DistillGPT
  - CheXRepair
  - MAIRA-2
- All report generators improved after corrected
- Radgraph F1 overestimates performance

Report Generator	Radgraph F1 (Automated)	Radgraph F1 (corrected)	Relative Improvement
RGRG[65]	0.52	0.67	28.8%
XrayGPT[66]	0.39	0.45	15.3%
GPT4o (inhouse)	0.43	0.51	18.6%
R2GenGPT[73]	0.54	0.58	7.4%
CV2DistillGPT2[45]	0.41	0.49	19.4%
CheXRepair[54]	0.38	0.43	13.1%
Maira-2[4]	0.58	0.63	8.6%

7-29% improvement in report quality through FC model!

FC Model as a surrogate ground truth

All 7 report generators evaluated against the test partitions of 4 datasets

Report Generator	CImaGenomeG *RQ*		MS-CXR *RQ*		ChestX-ray8 $\overline{FCScore}$		VinDR-CXR *RQ*	
	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)
RGRG[65]	0.541	0.537	0.329	0.308	0.305	0.298	0.549	0.537
XrayGPT[66]	0.622	0.626	0.388	0.391	0.377	0.355	0.618	0.609
GPT4-inhouse	0.658	0.653	0.433	0.426	0.399	0.408	0.636	0.630
R2GenGPT[73]	0.587	0.585	0.377	0.374	0.346	0.333	0.581	0.579
CV2DistillGPT2[45]	0.576	0.573	0.439	0.433	0.427	0.420	0.588	0.6
CheXRepair[54]	0.744	0.733	0.466	0.461	0.439	0.432	0.709	0.714
Maira-2[4]	0.619	0.633	0.423	0.425	0.412	0.419	0.578	0.569

Model	CCC (A,P)(A,G)
FCModel	0.997
FCSimple	0.831

FC model has high CCC!

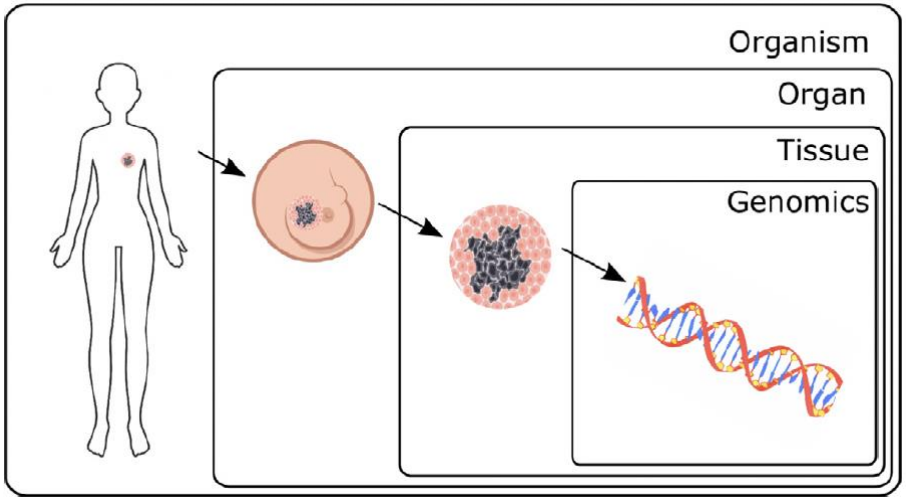
Architectures for multimodal fusion

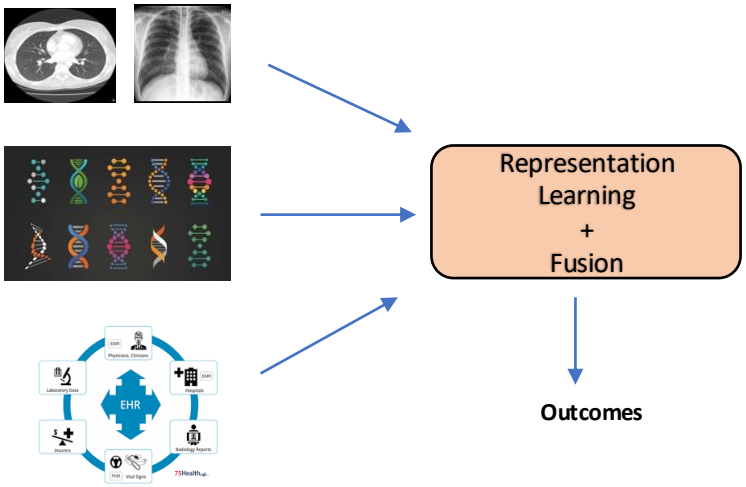
- Graph-based architectures taken from:

Fusing Modalities by Multiplexed Graph Neural Networks for Outcome Prediction in Tuberculosis, MICCAI'2022

MaxCorrMGNN: A Multi-Graph Neural Network Framework for Generalized Multimodal Fusion of Medical Data for Outcome Prediction, ICML Workshop on healthcare, 2023

Multimodal Fusion For Healthcare





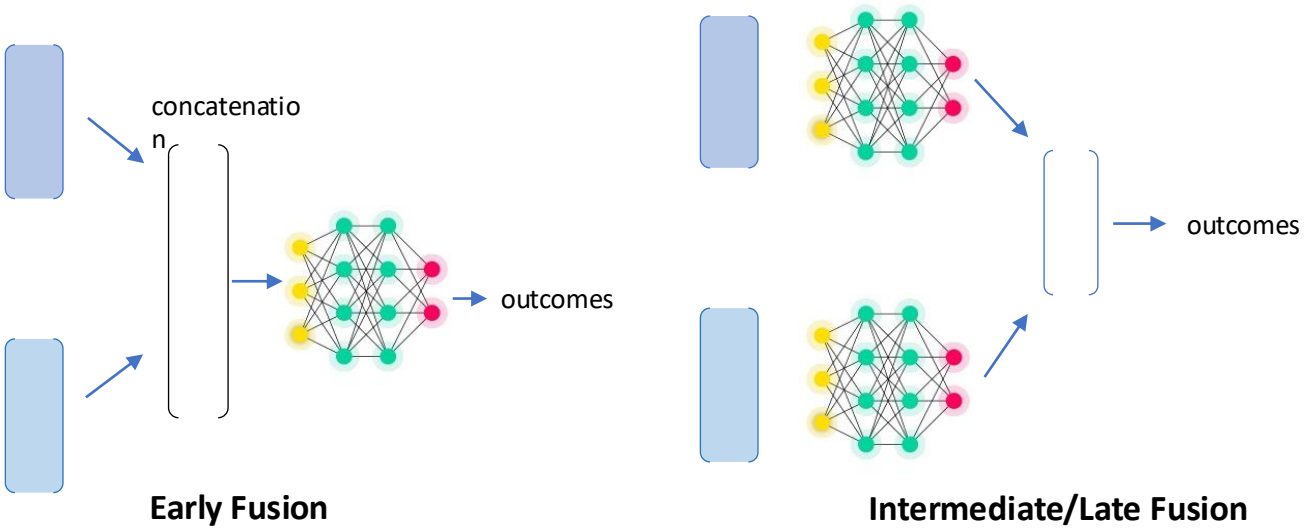
Challenges

- Inference from each modality is not definitive and may have uncertainty or errors
- Each modality may need a different representation to capture its information
- Not all modality evidence may be present or relevant for each instance of the problem
- Number of samples smaller than the dimensionality per modality

Multimodal fusion in healthcare - The challenges

- Representation
  - How to represent and summarize multimodal data across scales ?
  - Each modality may need a different representation to capture its information
- Alignment
  - How to capture the relationship between (sub)elements from two or more different modalities (spatial and temporal) across scales?
  - Capture modality relationships
- Fusion
  - How to combine information that is mutually reinforcing or complementary?
  - Fusion may need to be across both space and time
  - When does fusion help?
- Uncertainty handling
  - Inference from each modality is not definitive and may have uncertainty or errors
    - E.g. Evidence is pathology imaging assumes the tissue region was sampled correctly.
- Missing data
  - Not all modality evidence may be present for each instance of the problem
- Lack of coordinated datasets
  - No large-scale datasets that span all scales

Traditional Multimodal Fusion Strategies



Early Fusion **Intermediate/Late Fusion**

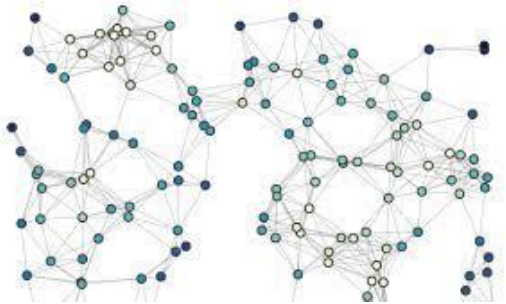
Limitations:

- Complexity grows with increasing number of concatenated features
- Learning and fusing representations at scales
- Exploiting correlations/inter-dependence within modality features

Modeling Multimodal Fusion through Graphs

- Multimodal data can be modeled as a graph, where each data object is regarded as a node, and both intra- and inter-modal dependencies existing between data objects can be regarded as edges.
- Aggregation of information through message passing for downstream task such as disease or outcome prediction
- Key problems:
  - Constructing the graph
    - As the graph structure is not part of a dataset of patient information, the crucial point of GNN-assisted disease prediction is the graph creation.
      - Graphs constructed through non-imaging features
      - Based on similarities in latent representations
      - Adaptive graph construction based on fused multimodal features, dynamically adjusted during model training.
      - Multimodal embeddings and the original feature values of patients through a weighted summation
      - Sparsity filtering
        - degree-sensitive edge pruning and kNN sparsification strategies to eliminate redundant and noisy edges during graph creation and merging.
  - Propagation through message passing
    - A form of representation learning
      - Allow granular interactions within and across different modalities to learn more robust data representations for multimodal
    - Graph interaction networks – model the relationships
    - Graph attention networks to capture intra-modal contextual information and inter-modal complementary information.
    - Graph isomorphism networks (GIN) – captures distinct graph structures

Graph Based Representation Learning



$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \ |\mathcal{V}| = P$$

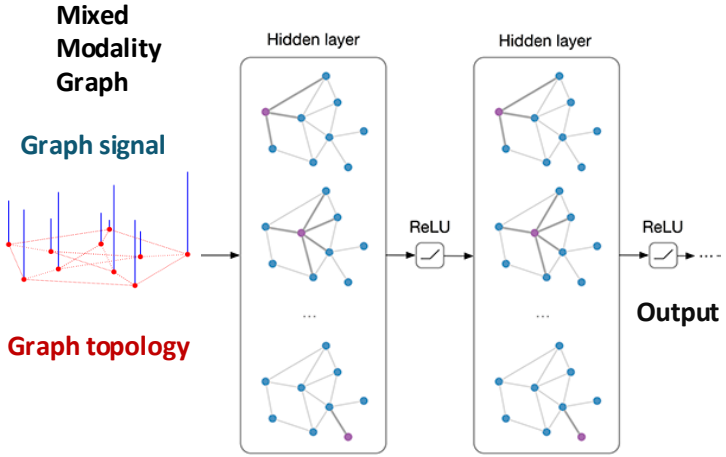
Nodes: Features/ Entities
Edges: Capture dependence within features

Adjacency Matrix: $\mathbf{A} \in \mathcal{R}^{P \times P}$

$$\text{If } (i,j) \in \mathcal{E}, \text{ then } \mathbf{A}(i,j) = 1$$
$$\text{else } \mathbf{A}(i,j) = 0$$

- Do not require spatial contiguity/organization of nodes
- Can represent data at arbitrary scales

Graph Neural Networks



Basic Message Passing

$$\mathbf{H}_i^{(l+1)} = \sigma\Big[\sum_{j \in N(i)} \mathbf{H}_j^{(l)} \mathbf{W}^{(l)}\Big]$$

non. linearity

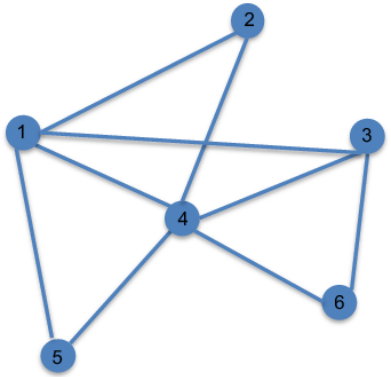
Messages at layer l

Filter at laye l

Graph topology: Intrinsic connectivity in the graph
Graph signal: Node/edge attributes for propagation

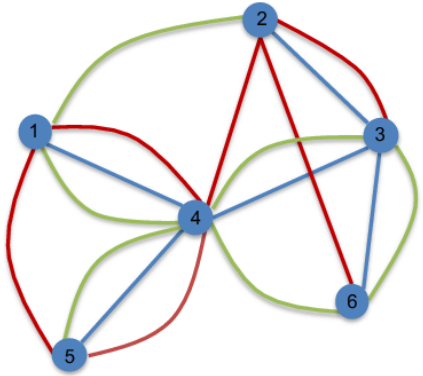
Message Passing: Tracking the information flow within the graph based on local neighbourhoods

Multi-Graphs for Multimodal Learning



Single view graph:
Graph with single edge types

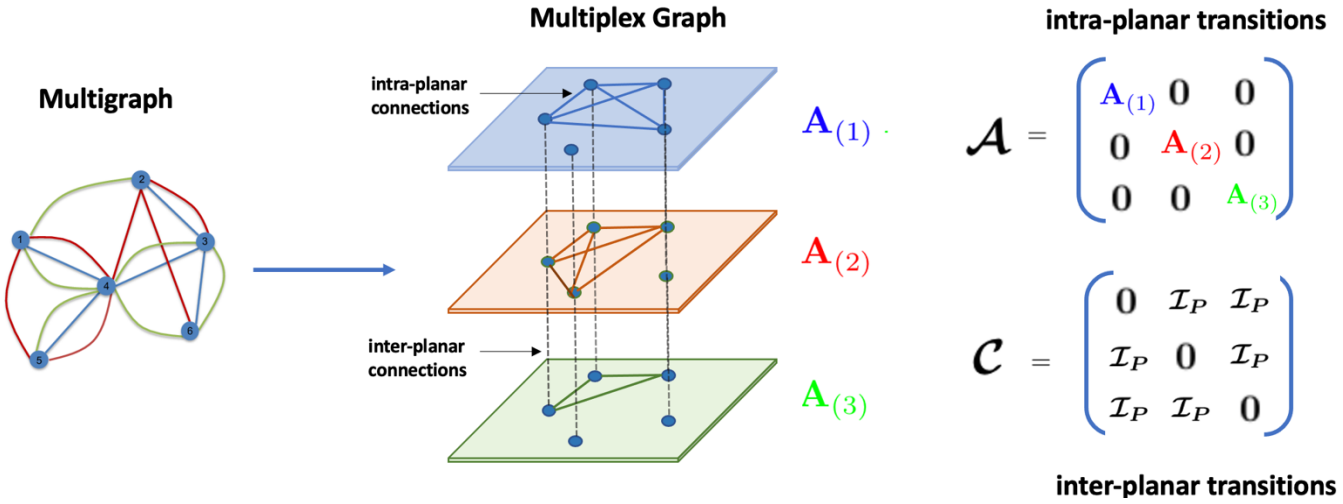
$$\mathbf{A} \in \mathcal{R}^{P \times P}$$



Mixed-modality Multi-graph:
Each colour is a distinct interaction pattern

$$\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)} \in \mathcal{R}^{P \times P}$$

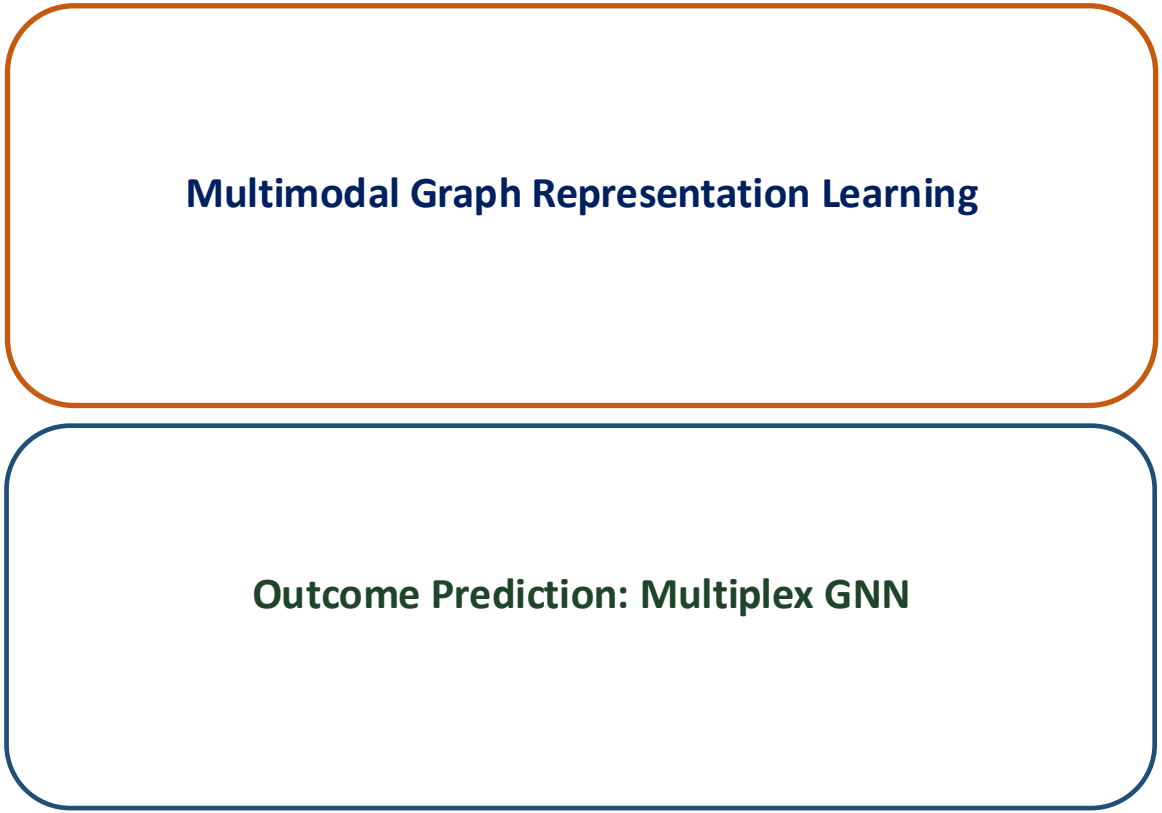
The Multiplex Framework



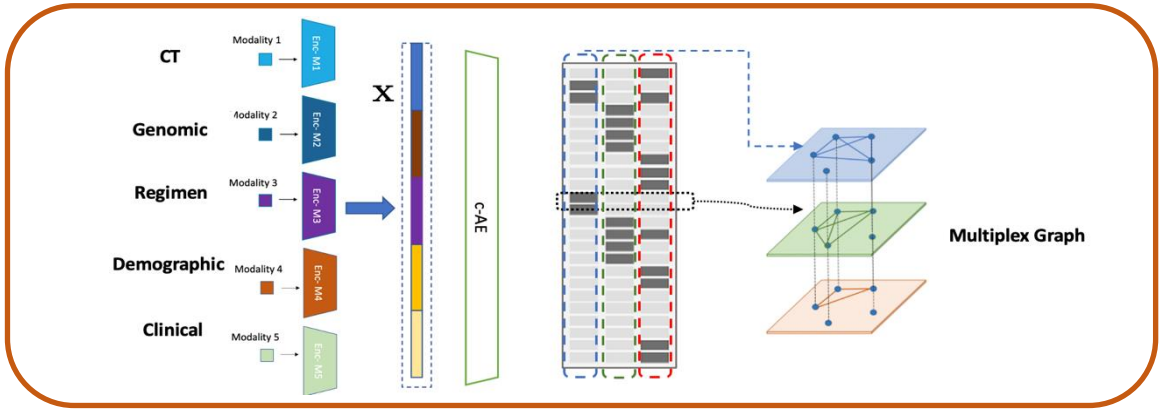
Motivation: Using intra- and inter "planar" connections to model different interaction patterns during multimodal fusion

Used earlier for transportation networks or social networks

Multimodal Fusion via Multiplex Graphs

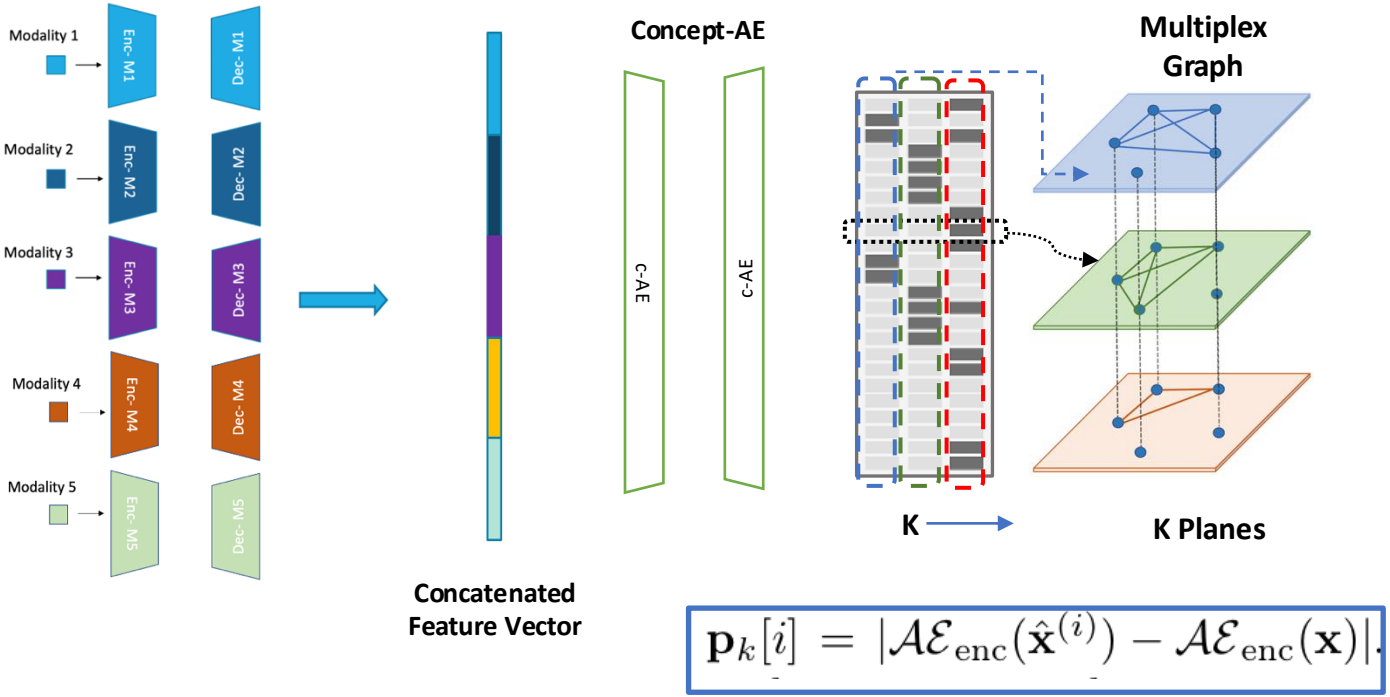


Multiplexed Graph Representation Learning



Outcome Prediction: Multiplex GNN

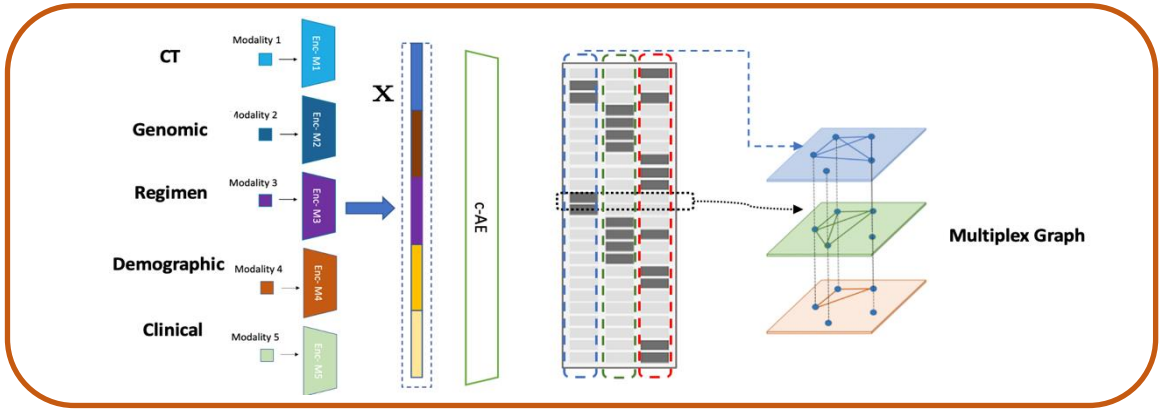
Multiplexed Graph Representation Learning

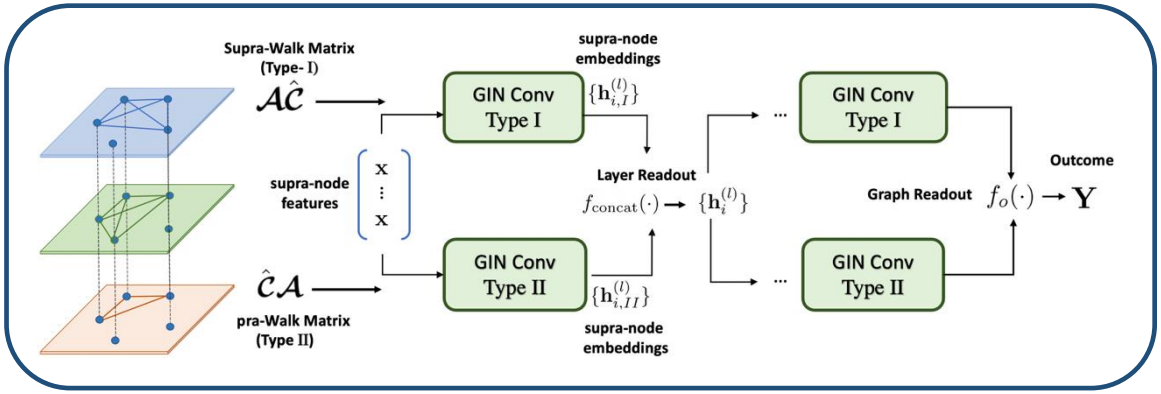


$$\mathbf{p}_k[i] = |\mathcal{AE}_{\text{enc}}(\hat{\mathbf{x}}^{(i)}) - \mathcal{AE}_{\text{enc}}(\mathbf{x})|.$$

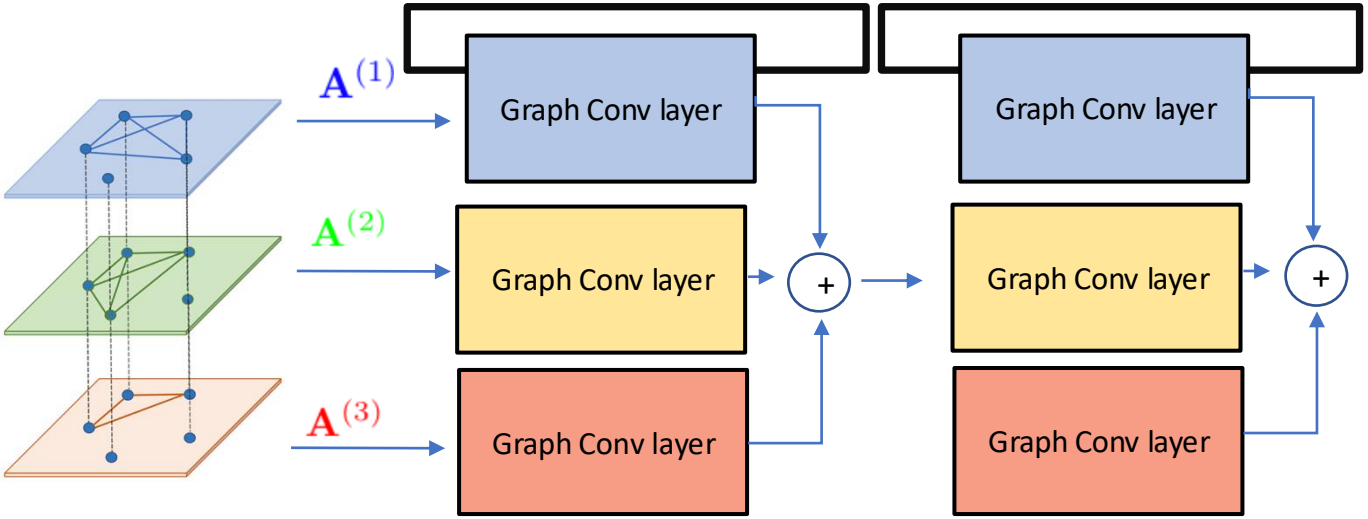
Each latent dimension of the autoencoder captures an abstract aspect of the interaction pattern

Multiplexed Graph Neural Networks





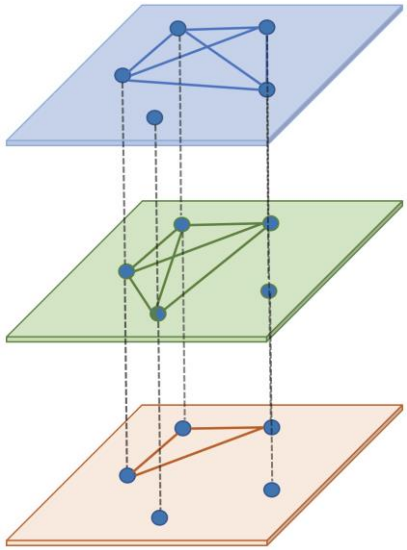
Prior Work on Multigraph Learning



Message Passing: $\mathbf{H}_i^{(l+1)} = \sigma\Big[\sum_{k=1}^{K}\sum_{j\in N(i)}\frac{1}{c_{i,k}}\mathbf{H}_j^{(l)}\mathbf{W}_k^{(l)} + \mathbf{H}_i^{(l)}\mathbf{W}_0\Big]$

- Separates message passing across edge-types and then aggregate
- May miss several higher order paths arising from cross modal connections

Defining Walks on the Multiplex



Intra-planar adjacency matrix

$$\mathcal{A} = \bigoplus_{k} \mathbf{A}^{(k)}$$

Inter-planar transition matrix

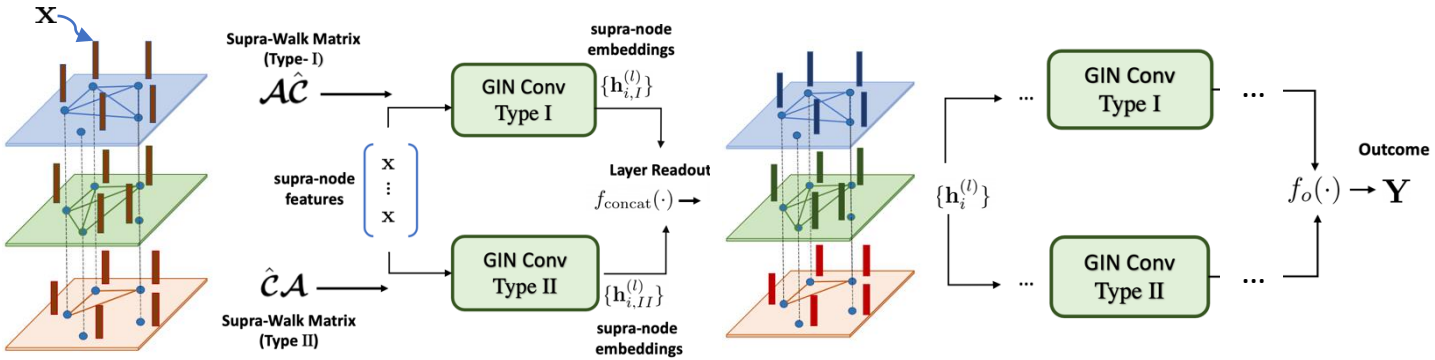
$$\mathcal{C} = [\mathbf{1}\mathbf{1}^T] \otimes \mathcal{I}_P - \mathcal{I}_{PK}$$

A walk consists of one of two types of steps:

I. First take intra-planar step, then allowed to transition $\mathcal{A}\hat{\mathcal{C}}$

II. First allowed to transition, then take intra planar $\hat{\mathcal{C}}\mathcal{A}$

Multiplex Graph Neural Network



GIN vs GCN

$$h_v^{(k)} = \text{MLP}^{(k)}\left((1+\epsilon)\cdot h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right)$$

$$\mathbf{h}_{i,I}^{(l+1)} = \text{GINConv}\left(\{\mathbf{h}_j^{(l)}, j : [\mathcal{A}\hat{\mathcal{C}}][i,j] = 1\}\right) \quad \text{Type I Filtering}$$

$$\mathbf{h}_{i,II}^{(l+1)} = \text{GINConv}\left(\{\mathbf{h}_j^{(l)}, j : [\hat{\mathcal{C}}\mathcal{A}][i,j] = 1\}\right) \quad \text{Type II Filtering}$$

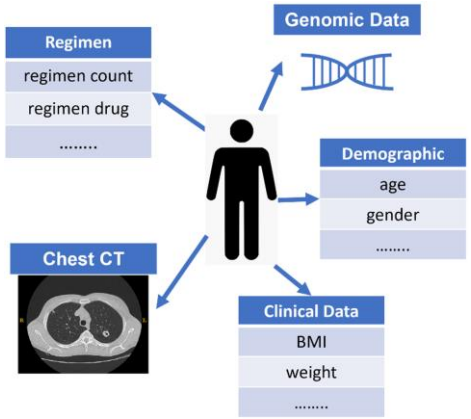
$$\mathbf{h}_i^{(l+1)} = f_{\text{concat}}(\mathbf{h}_{i,I}^{(l+1)}, \mathbf{h}_{i,II}^{(l+1)}) \quad \text{Layer-wise readout}$$

$$f_{\text{o}}(\{\mathbf{h}_i^{(L)}\}) = \mathbf{Y} \quad \text{Graph readout}$$

TB Dataset

Five modalities: [1 multiplex graph per patient]

- Imaging (CT annotations) – 2084 DenseNet features
- Genomic (SNP) - 4048 categorical features
- Demographic – 29 categorical, 1 continuous
- Clinical – 1726 categorical, 1 continuous
- Regimen – 233 categorical



Five Class Classification: [Graph level outcome] Cured, Failure, Still on treatment, Completed, Died

Baseline Comparisons

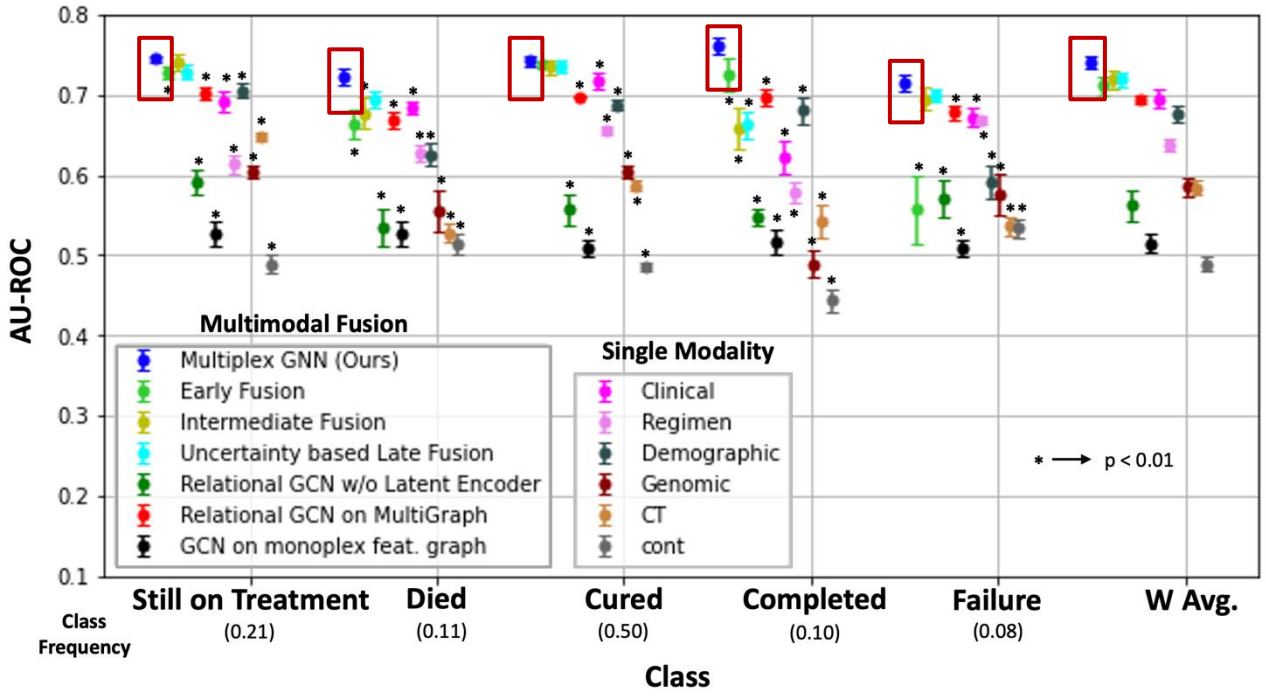
Fusion Baselines:

- No fusion /Individual modality prediction
- Early fusion
- Intermediate fusion
- Uncertainty based late fusion (Wang et. al. 2021)

Ablations:

- Relational GCN on a Multiplexed Graph (Schlichtkrull et. al. ,2018)
- Relational GCN w/o Latent Encoder (Schlichtkrull et. al. ,2018)
- GCN on monoplex feature graph (Kipf and Weling et. al., 2016)

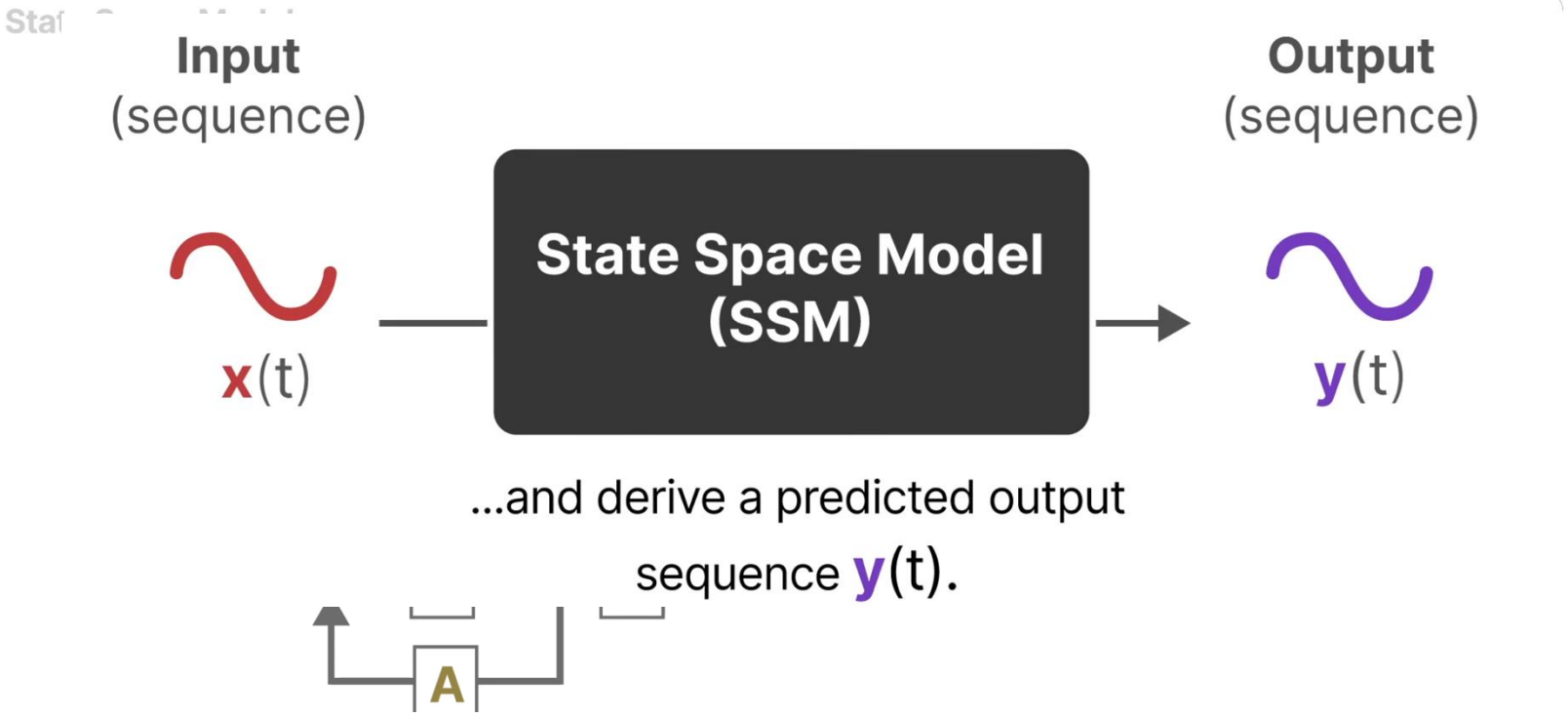
Results


AU-ROC
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Multimodal Fusion
Multiplex GNN (Ours)
Early Fusion
Intermediate Fusion
Uncertainty based Late Fusion
Relational GCN w/o Latent Encoder
Relational GCN on MultiGraph
GCN on monoplex feat. graph
Single Modality
Clinical
Regimen
Demographic
Genomic
CT
cont
* → p < 0.01
Still on Treatment
Died
Cured
Completed
Failure
W Avg.
Class Frequency
(0.21)
(0.11)
(0.50)
(0.10)
(0.08)
Class

Emerging architectures

- **State-space models**
- **Mamba**
- **Memory-augmented models/agents**
- Neuro-symbolic architectures
- Continuous learning foundation models
- Self-reflective / reasoning-optimized architectures
- Tool-integrated LLMs

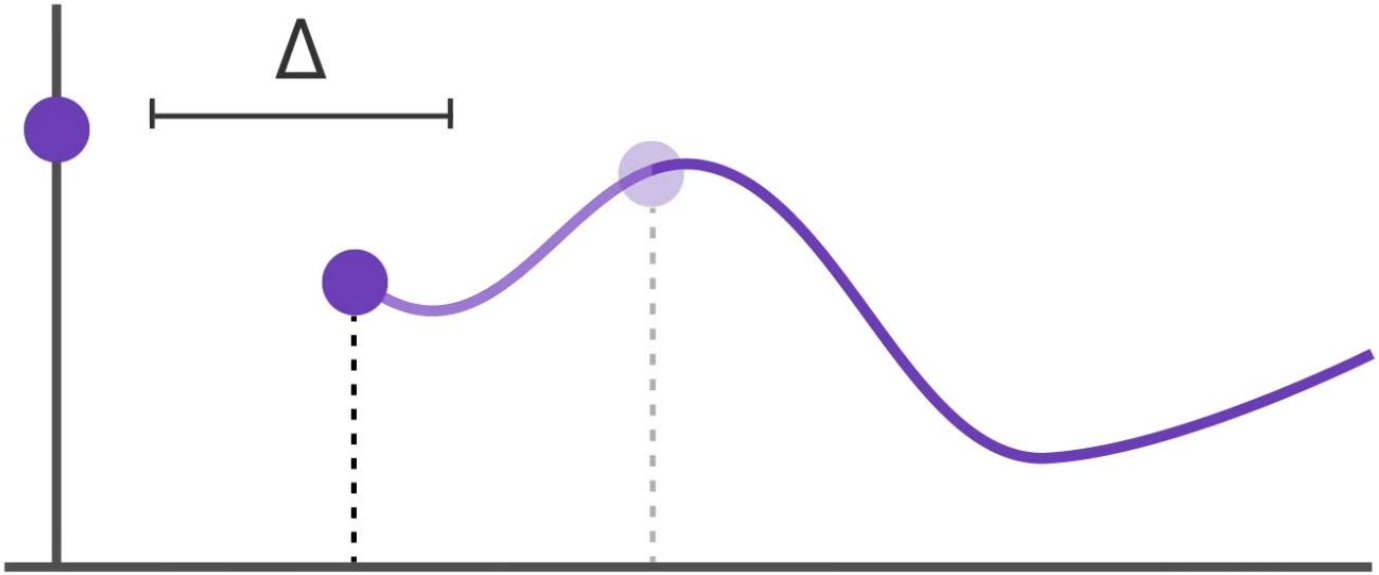
State-space models



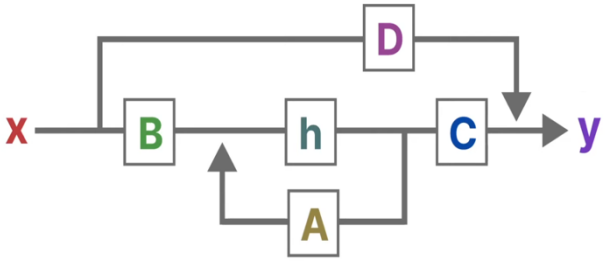
Source: Maarten Grootendorst

Efficiently Modeling Long Sequences with Structured State Spaces, ICLR 2022

Going from discrete to continuous state



LTI implies convolution for achieving parallelism



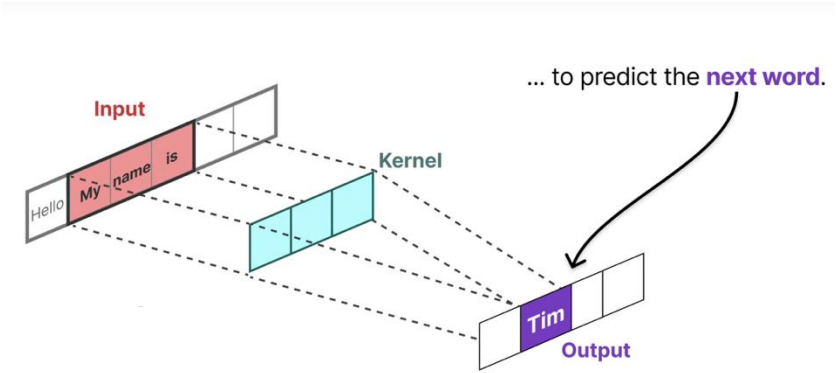
$$h_t = Ah_{t-1} + Bx_t$$

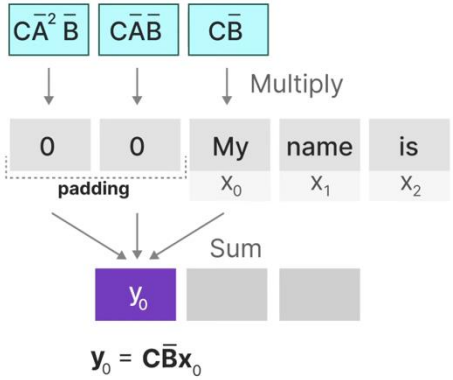
$$y_t = Ch_t$$

$$h_t = A^t h_0 + \sum_{k=0}^{t-1} A^k B x_{t-k}$$

$$K_k = CA^kB$$

Kernel



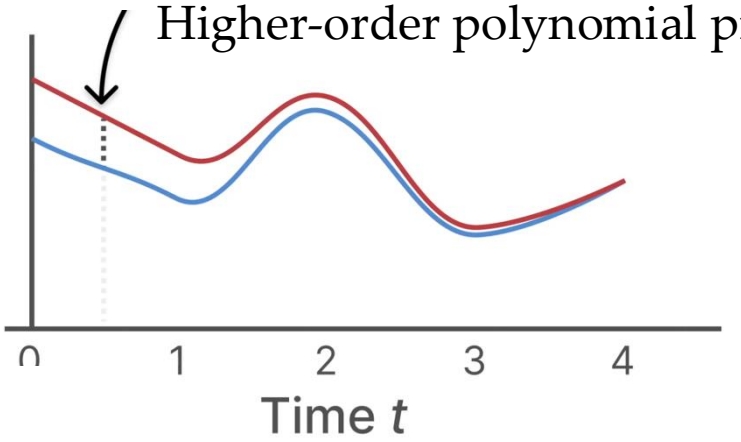


A^k

Controls how information persists. Eigenvalues close to 1-> long persistence

Compactly representing history in A matrix

- How should we choose A so that the hidden state stores as much useful information about the past as possible?
- A can be approximated by orthogonal polynomial projections and represented by their coefficients.
- Instead of remembering every past token, remember the best polynomial summary of the past.



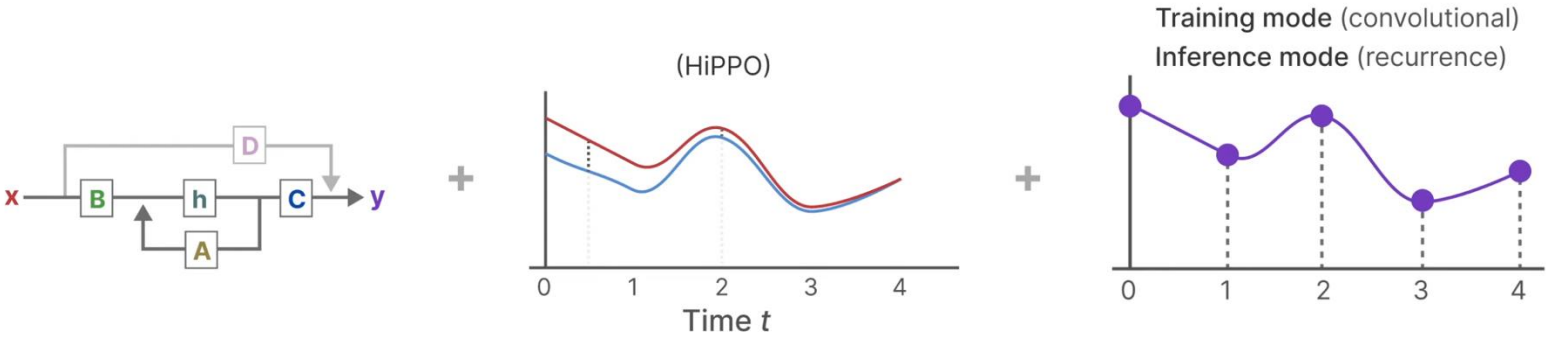
At time t, the hidden state represents:

$$\text{projection of } x(\tau),\ \tau \leq t$$

onto a set of orthogonal polynomials (e.g., Legendre polynomials).

https://proceedings.neurips.cc/paper/2020/file/102f0bb6efb3a6128a3c750dd16729be-Paper.pdf

Structured state space for sequences (S4) model



- A class of SSM that can efficiently handle long sequences:
  - Stays in continuous space but handles discrete input and output
  - Has a basic RNN structure
  - Combines the idea of convolution for the training for parallelism
  - Models history using orthogonal polynomial decomposition through progressively decaying older inputs.
- Still struggles with selective focusing for more refined language modeling
  - Obeying language constraints, e.g. for nouns

Mamba model

- Relaxes the fixed matrices to be dependent on input

$$h_t = Ah_{t-1} + Bx_t \longrightarrow h_t = A(x_t)h_{t-1} + B(x_t)x_t$$

System is no longer time-invariant

$$h_t = A^t h_0 + \sum_{k=0}^{t-1} A^k B x_{t-k}$$

$$h_t = A_t h_{t-1} + B_t x_t$$

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t$$

Linear in state-space

$$A_t = f_\theta(x_t)$$

A_t is diagonal (for efficiency)

Produced by a learned gating network

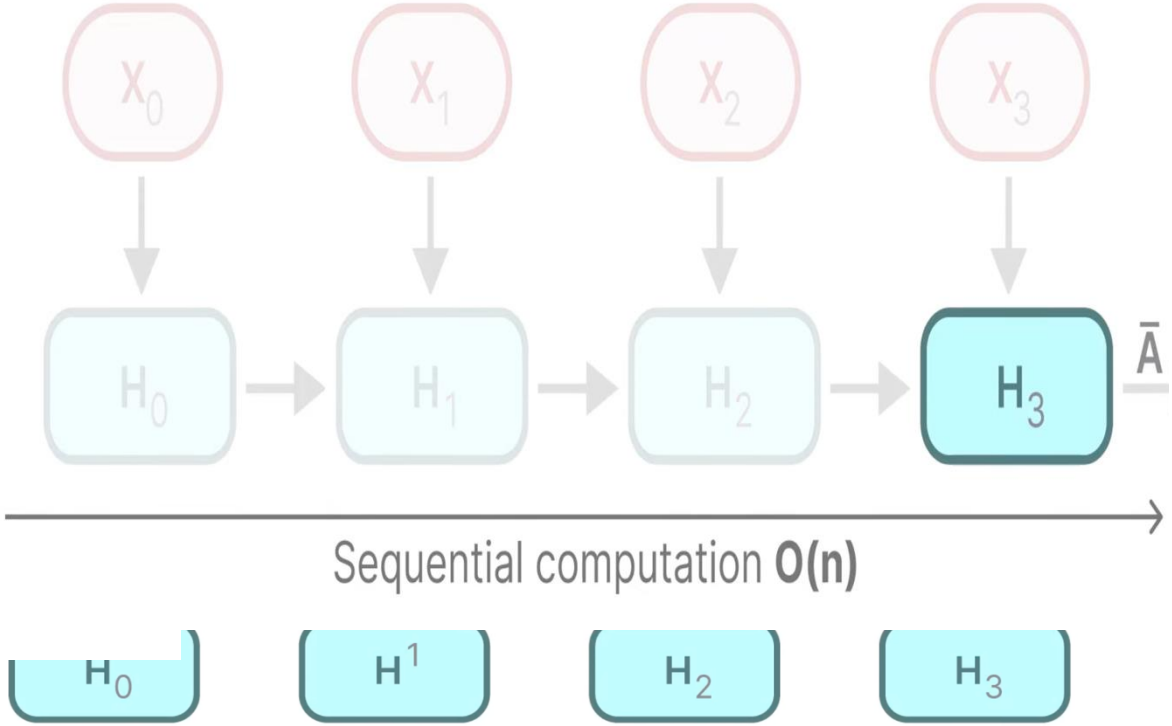
Varies per token

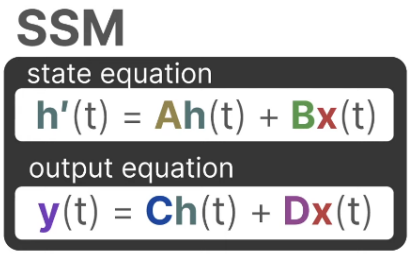
Unrolling, we see a sequence of matrix multiplications and prefix products which are both associative and hence parallelizable

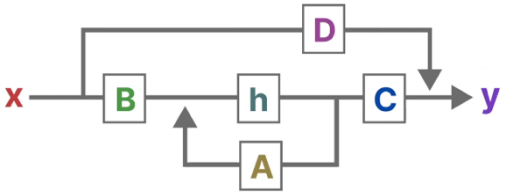
$$h_t = \bar{A}_t \bar{A}_{t-1} \dots \bar{A}_1 h_0 + \sum_{k=1}^{t} \left(\prod_{j=k+1}^{t} \bar{A}_j \right) \bar{B}_k$$

Mamba: Linear-Time Sequence Modeling with Selective State Spaces, COLM'2024

Parallel scan in Mamba







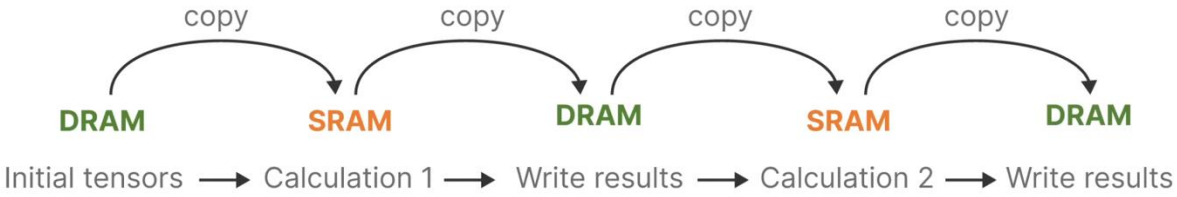
$a_1,\ a_1 \circ a_2,\ a_1 \circ a_2 \circ a_3,$

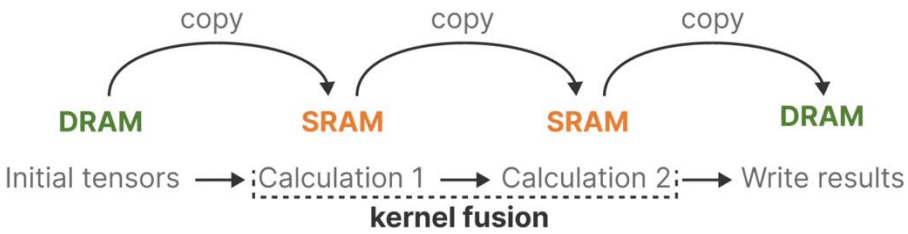
Can be computed in O(logn)

Hardware parallelism improvements in Mamba

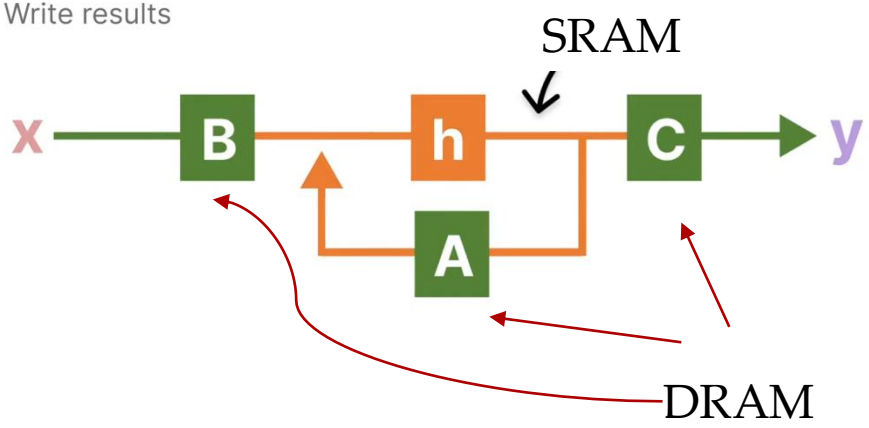
- SRAM-DRAM transfers can slow down GPUs

Selective scan + Hardware-awareness= Selective SSM (S6)

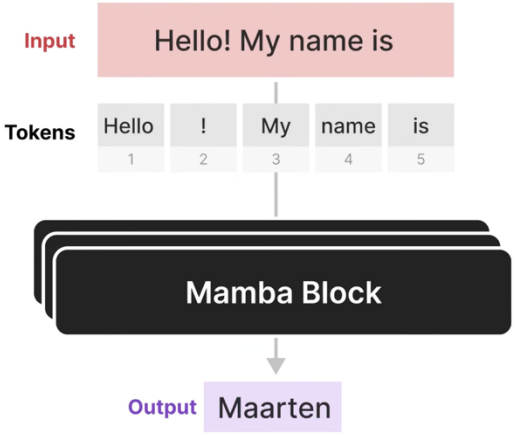


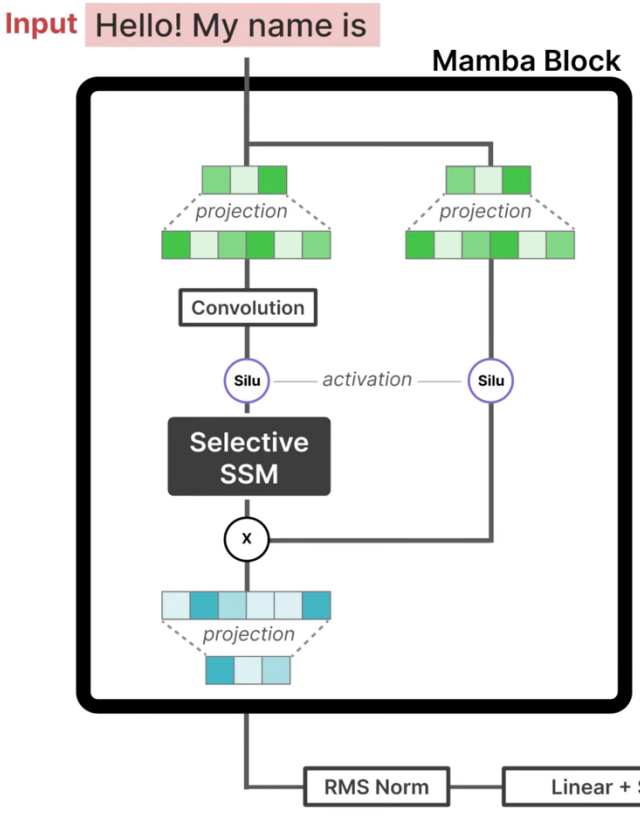


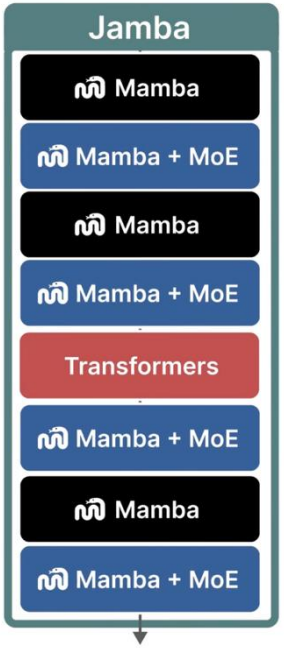
- Diagonal parameterization
- Vectorized state updates
- Associative scan
- Fused kernels
- Minimal memory movement



Mamba block







Can be intermixed with transformer blocks

Mamba Architecture

- S6 in a transformer-like block:
  - S4 model:
    - State-space model
    - Zero-order hold to go from discrete to continuous
    - HIPPO for memory modeling
    - Convolution during training
  - Selective scan algorithm (parallel scan + adaptive to input length in matrices)
  - Hardware aware for GPU parallelism
- Has over 8000 citations! So why isn't being adopted?

Memory-augmented networks

- **Memory-augmented networks** are neural models that include an **explicit external memory module inside the architecture**, trained end-to-end with the model.
- Memory is:
  - Differentiable
  - Learnable
  - Updated during training
  - Queried during forward passes
- Good for
  - Algorithmic reasoning
  - Storing intermediate results
  - Hard to scale
- Still experimental technology

- **Examples:**
  - Neural Turing Machine
  - Differentiable Neural Computer
  - Transformer variants with persistent memory slots
  - Memory layers in recurrent or state-space models

Larimar architecture

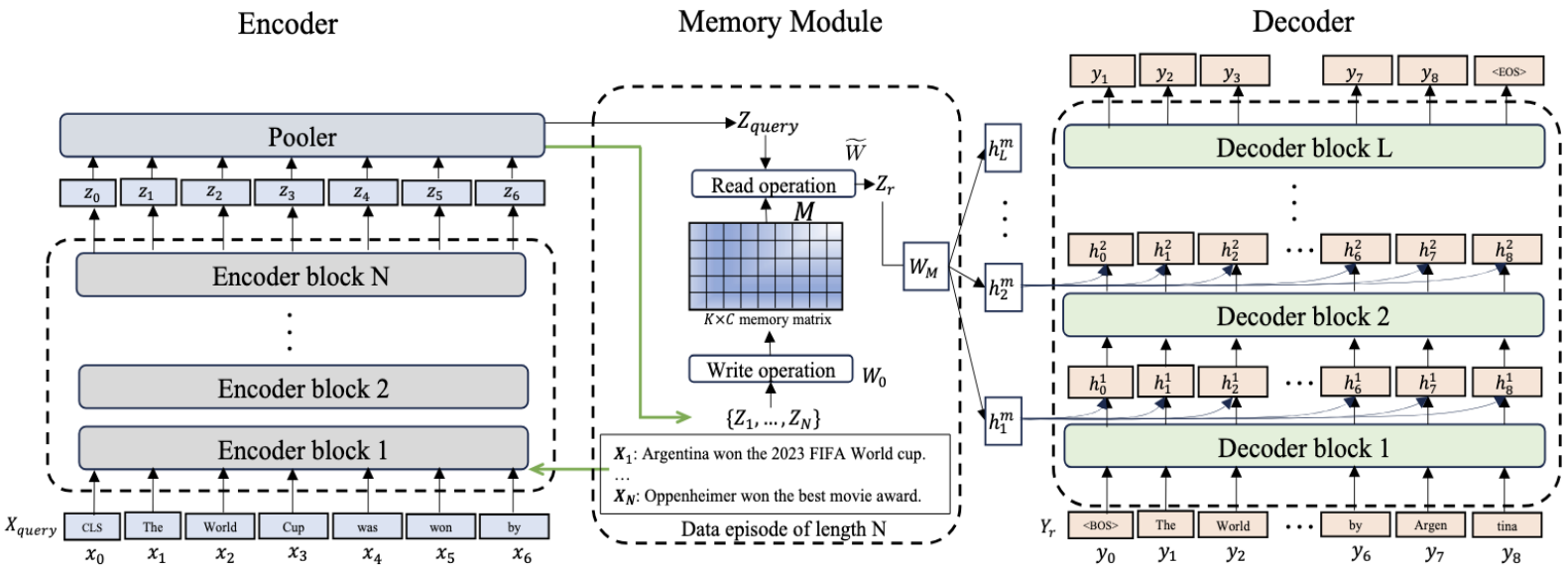


Figure 1. Larimar Architecture: X and X_{query} respectively denote data input and query, Z, Z_{query} and Z_r are the latent vectors, and M is the fixed-size memory. W and W_0 are reading/writing weights to memory. W_M interfaces the readout from memory to the decoder.

Larimar: Large Language Models with Episodic Memory Control, ICML'2024

Comparison of MAN (memory-augmented networks)

Architecture	Memory Type	Used During Training	Used During Inference	Memory Persistence	Notes
Larimar	Latent differentiable	✅	✅	Optional	Episodic, updatable facts
Transformer-XL	Cached hidden states	✅	✅	No	Extends context window
Compressive Transformer	Two-tier memory	✅	✅	No	Compress old memory
DNC / NTM	Explicit memory matrix	✅	✅	Optional	Algorithmic tasks
MPT-Memory	Latent memory buffer	✅	✅	Optional	Long-term conversation/ knowledge
RETRO	Retrieval database + latent memory	Partial	✅	Yes (DB)	Hybrid memory/retrieval

Most cited papers in FM (2026)

Paper	Citation	Paper	Citation	Paper	Citation
ResNet	309842	Attention	233902	Mask-RCNN	49109
LSTM	144352	BERT	162906	CLIP	21643
AlexNet	152817	U-Net	135230	Mamba	8815
Adam	237922	GPT-3	66603	FM	8760
ImageNet	152817	RAG	17372	SAM	18689

Top cited papers in FMs (2025)

- **ResNet:** 309842
- **LSTM :**
- **Attention Is All You Need,** *Vaswani et al. (2017),* 179562 citationsThis paper introduced the Transformer architecture, which revolutionized natural language processing by enabling parallel training and inference on long sequences of text.,
- •**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,** *Devlin et al. (2018),* 130,364This paper introduced BERT, a language model that uses bidirectional context to better understand the meaning of words in a sentence. BERT has become a widely used pretraining model in natural language processing.
- •**Convolutional Networks for Biomedical Image Segmentation, 110,730 (U-net)**
- •**Language Models are Few-Shot Learners,** *Brown et al. (2020),* 45399This paper introduced GPT-3, a language model that can perform a wide range of natural language tasks with little or no task-specific training. GPT-3 is notable for its large size (175 billion parameters) and its ability to generate coherent and convincing text.
- •**Mask R-CNN, 43054**
- •**Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,** *Radford et al. (2016)* (19,913)This paper introduced DCGANs, a type of generative model that uses convolutional neural networks to generate images with high fidelity.
- •**DALL-E: Creating Images from Text,** *Ramesh et al. (2021),*3307This paper introduced DALL-E, a generative model that can create images from textual descriptions. DALL-E has demonstrated impressive capabilities in generating realistic and imaginative images from natural language input.
- •**On the Opportunities and Risks of Foundation Models,** *Rishi Bommasani, Percy Liang, et al. (2021),*5420This paperhighlights progress made in the field of foundation models, while also acknowledging their risks—particularly the potential ethical and societal concerns, the impact on job displacement, and the potential for misuse by bad actors.