Lecture 10 - Contents

An overview of the parts in the multimodal medical AI lecture.

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 10:

Multimodal Medical AI: Integrating Text, Images, and Signals

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

Multimodal Medical Learning Overview

멀티모달 학습은 텍스트, 영상, 신호 등 **다양한 데이터 유형을 통합**하여 더 정확하고 포괄적인 의료 AI 모델을 구축

Text

임상 노트, 진단 보고서, EHR 데이터

Images

X-ray, CT, MRI, 병리 슬라이드

Signals

ECG, EEG, 생체 신호, 파형

Genomics

DNA/RNA 시퀀스, 유전자 변이

Video

수술 영상, 내시경, 동작 분석

Audio

폐음, 심음, 청진 데이터

통합 진단

시너지 효과

단일 모달리티 대비 향상된 성능

임상 실용성

실제 의료 환경과 유사한 통합 접근

다양한 데이터 소스로부터 종합적인 진단

Part 1:

Medical Vision-Language Models

Medical Image Encoders

CNN (Convolutional Neural Networks)

전통적이지만 강력한 영상 특징 추출 방법

- 공간적 계층 구조 학습
- 로컬 패턴 인식에 강점
- ResNet, DenseNet, EfficientNet
- 의료 영상의 텍스처, 경계 검출

Vision Transformer (ViT)

패치 기반 어텐션으로 글로벌 관계 학습

- Self-attention 메커니즘
- 긴 범위 의존성 포착
- 대규모 데이터셋에서 우수
- 병변 간 관계, 전체 맥락 이해

의료 특화 인코더

의료 도메인에 사전 학습된 모델

- MedCLIP, BiomedCLIP
- 대규모 의료 영상-텍스트 쌍 학습

Hybrid Architectures

CNN과 Transformer의 결합

- CNN: 로컬 특징 + ViT: 글로벌 관계
- CoAtNet, MaxViT

- 도메인 지식 내재화
- 적은 데이터로 높은 성능

- 계산 효율성과 성능 균형
- 다양한 스케일의 병변 검출

의료 영상 전처리 파이프라인

정규화 (Normalization) → 윈도잉 (Windowing) → 리샘플링 (Resampling) → 증강 (Augmentation)

특징 추출

고차원 영상을 저차원 벡터 표현으로 변환하여 의미 있는 임상 특징 포착

다운스트림 태스크

추출된 특징을 분류, 검출, 세분화, 리포트 생성 등에 활용

CLIP Medical Variants: Contrastive Learning

CLIP (Contrastive Language-Image Pre-training): Aligns images and text in a common embedding space, enabling **zero-shot learning**

Common Embedding Space

Image Encoder

ResNet / ViT

Image feature extraction

→

Projection

Common

Embedding space

←

Text Encoder

Transformer

Text feature extraction

MedCLIP

Trained on medical image-text pairs, general disease diagnosis

BiomedCLIP

15 million biomedical image-text pairs, broad domain coverage

PubMedCLIP

Based on PubMed papers, scientific image understanding

RadCLIP

Specialized in radiology images, X-ray/CT reports

PathCLIP

Dedicated to pathology slides, tissue classification

CXR-CLIP

Focused on chest X-rays, lung disease detection

Contrastive Learning

- Matching image-text pairs: high similarity

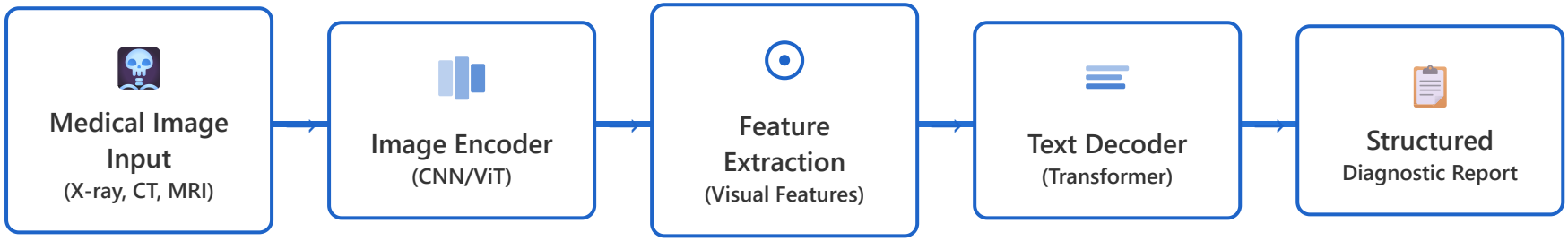
Image-Text Alignment Effects

- Enables zero-shot classification

- Non-matching pairs: low similarity
- Discriminative learning with InfoNCE loss

- Image retrieval with text queries
- Multimodal representation learning

Radiology Report Generation: Automated Medical Report Writing



Chest X-ray

Automated description of chest diseases like pneumonia, tuberculosis, pleural effusion

CT Scan

Detailed description of lesion location, size, and characteristics by organ

MRI

Text generation for abnormal findings in brain, spine, and joints

Mammography

Report interpretation of breast calcification and masses

Ultrasound

Automated documentation of ultrasound findings

Multi-View

Integrated analysis of multiple angle images

Structured Report Sections


Findings
Impression
Comparison
Recommendation

Pathology Slide Analysis: WSI Analysis

WSI (Whole Slide Imaging): Gigapixel-level pathology slides **divided into patches** for tissue classification and lesion detection

1. Slide Scanning

- High-resolution digital scanning (20x, 40x)
- Images of tens of GB in size
- Multi-resolution pyramid structure

2. Patch Extraction

- 256x256 or 512x512 patch division
- Tissue region detection (background removal)
- Overlap setting via stride adjustment

3. Patch-level Analysis

- Feature extraction per patch using CNN/ViT
- Tissue type classification (normal/cancer/inflammation, etc.)
- Multiple Instance Learning (MIL)

4. Slide-level Integration

- Patch prediction aggregation (mean, max, attention)
- Whole slide diagnosis (cancer grade, prognosis)
- Heatmap visualization

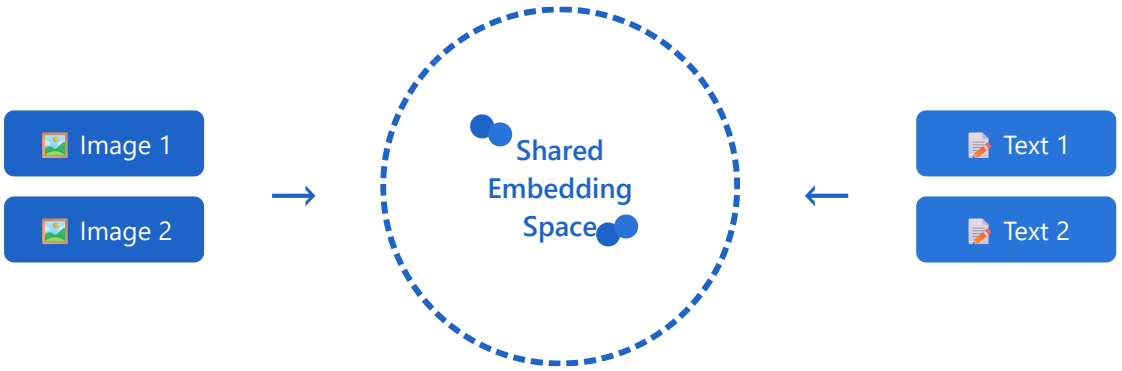
Cancer Detection & Grade Classification

Lymph Node Metastasis Detection

Ki-67 Proliferation Index Measurement

Cross-Modal Alignment: Inter-Modal Alignment

By placing different modalities at semantically close positions in a shared embedding space, mutual understanding and retrieval become possible



Projection Layers

Linear/non-linear transformations that project each modality into a shared space of the same dimension

- Dimension matching with linear layers
- Non-linear transformation with MLP
- Placement on unit sphere via L2 normalization

Contrastive Loss

Improves alignment quality by learning to bring matching pairs closer and push non-matching pairs apart

- Using InfoNCE Loss
- Applying temperature scaling
- Hard negative mining

Triplet Loss

Cross-Attention

Learning relative distances with Anchor, Positive, and Negative samples

- Same patient data: Positive
- Different patient data: Negative
- Margin-based distance optimization

Enhancing inter-modal interaction through Transformer-based attention

- Query-Key-Value mechanism
- Information exchange between modalities
- Dynamic weight learning

Zero-shot Classification

Text→Image Retrieval

Image→Text Retrieval

Contrastive Learning

Learning to distinguish positive pairs (different views of the same patient, same diagnosis) and negative pairs to improve medical representation quality

Positive Pairs

Semantically similar data pairs

- Images of the same patient at different timepoints
- Different modalities of the same disease
- Image-report matching pairs
- Data augmented same samples

Negative Pairs

Semantically different data pairs

- Images from different patients
- Different disease categories
- Non-matching image-report pairs
- Different samples within batch

InfoNCE Loss

Core loss function for contrastive learning

- Positive pairs: High similarity (cosine)
- Negative pairs: Low similarity
- Temperature parameter adjustment
- Requires large batch size

Medical Domain Application

Medical-specialized contrastive learning strategies

- Leveraging anatomical consistency
- Preserving temporal continuity
- Integrating clinical metadata
- Pre-training without labels

Self-supervised

Few-shot

Enhanced

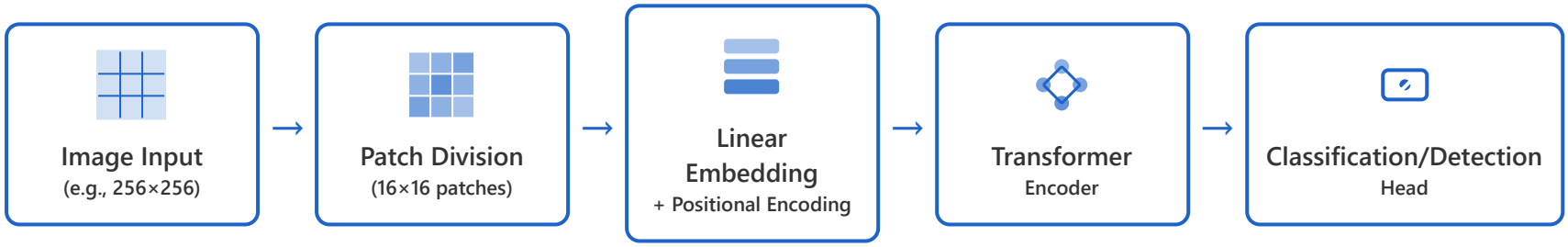
Pre-training

Learning Enabled

Transfer Learning

Vision Transformer (ViT) for Medical Imaging

Process medical images with patch embedding and positional encoding, capture global relationships through self-attention



Self-Attention Mechanism

Learning long-range dependencies across entire image

- Compute relationships between all patches
- Query, Key, Value transformations
- Capture spatial relationships between lesions
- Multi-head attention

Medical Imaging Applications

ViT specialized for medical domain

- High-resolution image processing
- Variable-size input support
- 3D volume extension (3D-ViT)
- Medical-specific pre-training

Advantages

Strengths of ViT

- Global context understanding

Medical ViT Models

Representative medical-specialized ViT

- MedViT: Medical image pre-training

- Scalable architecture
- Large-scale pre-training benefits
- Various downstream tasks

- TransUNet: Segmentation specialized
- Swin Transformer: Hierarchical
- CoTr: CT/MRI reconstruction

Part 2:

Beyond Images - Signals and Sequences

Audio: Lung Sounds Analysis

Classify respiratory diseases such as pneumonia, asthma, and COPD using CNN after converting auscultation sounds to spectrograms

Types of Lung Sounds

Major auscultation sound patterns

- Normal lung sounds: Vesicular sounds
- Wheezes: Airway constriction
- Crackles: Fluid/inflammation
- Rubs: Pleural inflammation

Audio Preprocessing

Acoustic signal transformation and augmentation

- Spectrogram generation (STFT)
- Mel-frequency transformation
- Noise removal (heart sounds, etc.)
- Time segmentation (windowing)

Deep Learning Architecture

Acoustic classification models

- 2D CNN: Spectrogram processing
- ResNet: Deep feature extraction
- RNN: Temporal patterns
- Audio Transformer: Attention

Multimodal Fusion

Integration of auscultation sounds with other data

- Lung sounds + Chest X-ray
- Lung sounds + Pulmonary function tests
- Lung sounds + Clinical symptoms
- Spatial-acoustic alignment

Early pneumonia detection

COPD monitoring

Asthma exacerbation prediction

Audio: Lung Sounds Analysis

청진음을 스펙트로그램으로 변환 후 CNN으로 폐렴, 천식, COPD 등 호흡기 질환 분류

폐음 유형

주요 청진음 패턴

- 정상 폐음: Vesicular sounds
- 천명음 (Wheezes): 기도 협착
- 수포음 (Crackles): 수분/염증
- 마찰음 (Rubs): 흉막 염증

오디오 전처리

음향 신호 변환 및 증강

- 스펙트로그램 생성 (STFT)
- Mel-frequency 변환
- 노이즈 제거 (심장음 등)
- 시간 분할 (윈도우)

딥러닝 아키텍처

음향 분류 모델

- 2D CNN: 스펙트로그램 처리
- ResNet: 깊은 특징 추출
- RNN: 시간적 패턴
- Audio Transformer: 어텐션

멀티모달 융합

청진음과 다른 데이터 통합

- 폐음 + 흉부 X-ray
- 폐음 + 폐기능 검사
- 폐음 + 임상 증상
- 공간-음향 정렬

폐렴 조기 발견

COPD 모니터링

천식 악화 예측

Genomic Sequence Integration

Processing DNA/RNA sequences with Transformer or CNN, integrated with mutation detection and phenotype prediction

Genomic Data Types

Genetic information used in medical AI

- DNA sequences (A, T, G, C)
- RNA expression profiles
- Mutations (SNP, CNV, Indel)
- Epigenetics (methylation)

Sequence Encoding

Converting base sequences to numerical representations

- One-hot encoding
- K-mer embedding
- Positional encoding
- DNA-BERT, DNA2Vec

Deep Learning Models

Neural networks for genome analysis

- 1D CNN: motif detection
- Transformer: long-range patterns
- Graph NN: gene networks
- VAE: latent representation learning

Multi-omics Integration

Genomic data with other modalities

- Genomics + Imaging (Radiophenomics)
- Genomics + Pathology (Pathogenomics)
- Genomics + Clinical records
- Multi-view learning

Cancer gene	Personalized treatment	Drug response

mutation detection

prediction

assessment

Time Series: Vital Signs Monitoring

Modeling ICU vital signals (HR, BP, SpO2) with RNN/LSTM/Transformer for deterioration prediction

Key Vital Signs

Monitored vital signs

- Heart Rate (HR): Beats per minute
- Blood Pressure (BP): Systolic/Diastolic
- Oxygen Saturation (SpO2): Blood oxygen
- Respiratory Rate (RR): Breaths per minute

Time Series Characteristics

Temporal patterns of vital signals

- Irregular sampling intervals
- Missing values (sensor errors)
- Trends and periodicity
- Abrupt changes (events)

Deep Learning Models

Time series prediction architectures

- LSTM: Long-term dependencies
- GRU: Lightweight recurrent network
- Temporal CNN: Dilated convolutions
- Transformer: Attention-based

Multimodal Integration

Vital signals and other data

- Vitals + Laboratory tests
- Vitals + Medication records
- Vitals + Clinical notes
- Time alignment required

Early Sepsis Warning

Cardiac Arrest Risk Prediction

ICU Length of Stay Estimation

3D Medical Imaging

Processing CT/MRI volume data with 3D CNN, utilizing spatial context for lesion detection and segmentation

3D Imaging Modalities

Volume Data Sources

- CT: Continuous slices (axial)
- MRI: Multi-sequence (T1, T2, FLAIR)
- PET/SPECT: Functional imaging
- Ultrasound: 3D reconstruction

3D CNN Architecture

Networks for volume processing

- 3D convolution kernels (3×3×3)
- 3D U-Net: Segmentation
- V-Net: Medical specialization
- nnU-Net: Automatic configuration

Spatial Features

Advantages of 3D

- Utilizing Z-axis context information
- Volume measurement accuracy
- 3D structure preservation
- Adjacent slice correlation

Multimodal 3D Integration

Multiple volume data fusion

- CT + PET: Anatomy + Function
- MRI multi-sequence fusion
- 3D imaging + Clinical data
- Spatial registration

Lung Nodule Detection

Brain Tumor Segmentation

Organ Volume Measurement

Video: Surgical Analysis

Frame-by-frame analysis of surgical videos for phase recognition, tool tracking, and complication prediction

Surgical Video Characteristics

Unique properties of video data

- High resolution (HD, 4K)
- Long duration (1-6 hours)
- Complex scene changes
- Various surgical tools

Video Processing Techniques

Spatiotemporal information extraction

- Frame sampling (FPS adjustment)
- 2D CNN + RNN/LSTM
- 3D CNN: spatiotemporal convolution
- Optical flow: motion analysis

Surgical Phase Recognition

Surgical workflow analysis

- Phase recognition (7-10 phases)
- Action segmentation
- Temporal CNN (TCN)
- Real-time feedback

Multimodal Integration

Video and other data sources

- Video + kinetic data
- Video + surgical records
- Video + patient information
- Time synchronization critical

Tool tracking and

Surgical skill

Complication risk

usage analysis

objective assessment

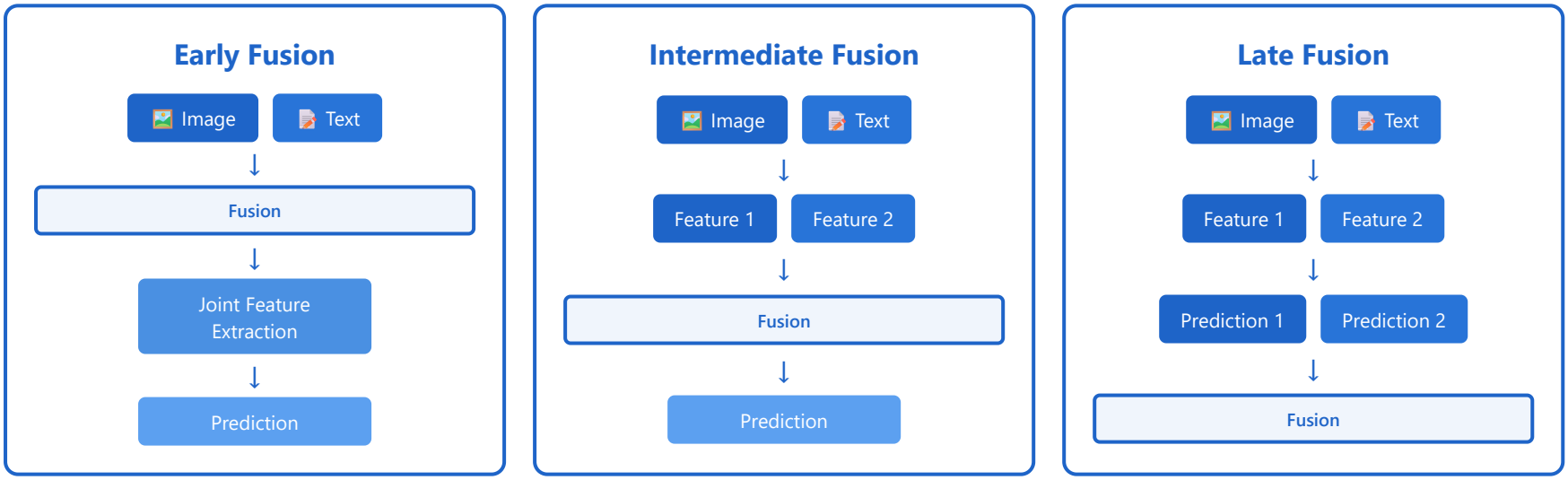
early detection

Part 3:

Multimodal Fusion Techniques

Early vs Late Fusion: Fusion Timing Strategies

Early Fusion: Combine before feature extraction / Late Fusion: Integrate after independent processing of each modality



Early Fusion

Input-level combination

- Combine raw data or early features
- Joint learning with single model

Pros: Strong integration

Cons: Curse of dimensionality

Intermediate Fusion

Mid-level combination

- Combine after partial processing of each modality
- Integration at intermediate representation level

Pros: Balanced approach

Cons: Design complexity

Late Fusion

Decision-level combination

- Complete independent processing of each modality
- Combine high-level predictions

Pros: Modality independence

Cons: Weak integration

Medical Application Examples

Early: CT + PET pixel-level fusion

Intermediate: X-ray + text feature combination

Late: Integration of imaging/genomic/clinical predictions

Selection Criteria

Early: Similar modality size/format, closely related

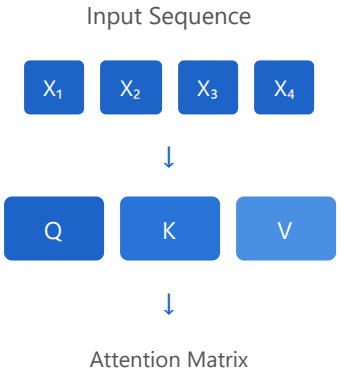
Late: Heterogeneous modalities, missing data possible

Hybrid: Complex multimodal systems

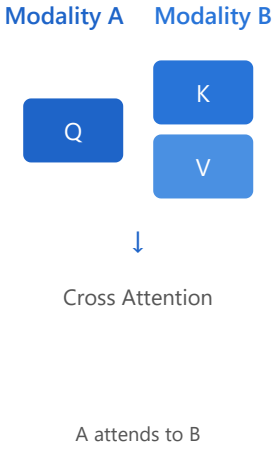
Attention-Based Fusion

Fusion by dynamically adjusting inter-modal importance using Cross-attention and Self-attention

Self-Attention



Cross-Attention



Self-Attention

Internal relationships within a single modality

- Query, Key, Value transformation
- Calculate similarity between all positions
- Update representation with weighted sum

Cross-Attention

Inter-modal interaction

- Query from one modality
- Key, Value from another modality
- Information exchange between modalities

- Capture long-range dependencies

- Selective information extraction

Multi-Head Attention

Fusion from multiple perspectives

- Multiple attention heads in parallel
- Each head focuses on different features
- Concatenate results
- Enhanced representation power

Medical Applications

Attention-based medical fusion

- Image-to-text attention
- Multi-timepoint image fusion
- Time-series + static data
- Interpretable weights

Image-Report Alignment

Multi-Sequence MRI Fusion

Clinical Data Integration

Graph Neural Networks (GNN) for Multimodal Fusion

Improving fusion quality by modeling inter-modal relationships with medical knowledge graphs and GNN

Medical Knowledge Graph

Structured medical knowledge representation

- Nodes: Diseases, symptoms, tests, treatments
- Edges: Relationships (causes, diagnosis, treatment)
- Attributes: Probability, frequency, severity
- Ontology integration (SNOMED, ICD)

GNN Architecture

Graph-based learning

- Graph Convolution (GCN)
- Graph Attention (GAT)
- Message Passing
- Graph Pooling

Multimodal Graph Construction

Integrating heterogeneous data into graphs

- Each modality = Node type
- Inter-modal relationships = Edges
- Patient-specific subgraphs
- Dynamic graph updates

Medical Applications

GNN-based fusion cases

- Disease prediction: Symptom graphs
- Drug interaction modeling
- Patient similarity graphs
- Knowledge-based reasoning

Multiple Disease Comorbidity Prediction

Personalized Treatment Recommendation

Clinical Pathway Optimization

Hierarchical Integration

Hierarchical integration from low-level features to high-level semantics through multi-stage fusion

Level 1 — **Low-level Feature Fusion**

Raw data or initial features: Texture, edges, pixel intensity, frequency components

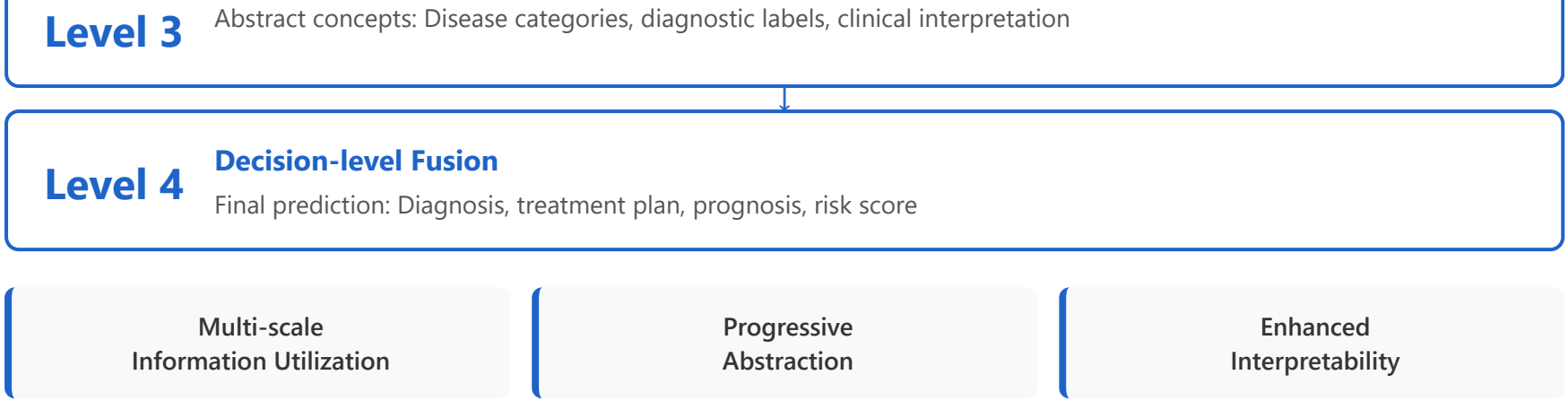
↓

Level 2 — **Mid-level Feature Fusion**

Partial semantic representation: Anatomical structures, object parts, pattern combinations

↓

High-level Semantic Fusion


Level 3
Abstract concepts: Disease categories, diagnostic labels, clinical interpretation
Level 4
Decision-level Fusion
Final prediction: Diagnosis, treatment plan, prognosis, risk score
Multi-scale
Information Utilization
Progressive
Abstraction
Enhanced
Interpretability

Missing Modality Handling

Applying alternative strategies (zero-padding, imputation, robust fusion) when some modalities are absent

Zero-padding / Masking

Simple baseline strategy

- Missing modality: Fill with zeros
- Use mask tokens
- Easy implementation, fast inference
- Possible performance degradation

Imputation

Predict and fill missing data

- Mean/median imputation
- K-NN based imputation
- Generative models (VAE, GAN)
- Cross-modal prediction

Robust Fusion

Design fusion robust to missing data

- Dropout-based training
- Leverage late fusion
- Ensemble approach
- Gating mechanism

Knowledge Distillation

Transfer knowledge from complete model

- Teacher: Uses all modalities
- Student: Handles partial modalities
- Soft label learning
- Prepare for missing scenarios

Random dropout

Modality-specific

Dynamic

during training

confidence weighting

architecture adjustment

Clinical Decision Support: Clinical Decision Support Systems

Multimodal CDSS provides integrated diagnosis, treatment recommendations, and prognosis prediction

Diagnostic Support

Multi-data based diagnosis

- Imaging + Laboratory tests
- Symptoms + Medical history
- Vital signs + Physical examination
- Probabilistic diagnosis presentation

Treatment Recommendations

Personalized treatment planning

- Drug selection and dosage
- Surgery vs Conservative treatment
- Side effect risk assessment
- Guideline compliance

Prognosis Prediction

Long-term outcome estimation

- Survival rate prediction
- Complication risk level
- Recurrence probability
- Functional recovery expectation

Real-time Monitoring

Continuous patient observation

- Early deterioration warning
- Automated alert system
- Trend analysis
- Intervention timing suggestion

Emergency Room Triage

ICU Management

Chronic Disease Management

Performance Benchmarks: Performance Evaluation

Comparing multimodal models vs. single-modal benchmarks and evaluation metrics (accuracy, AUC, F1)

Accuracy

Overall accuracy
Ratio of correct predictions

AUC-ROC

Classification performance
Sensitivity-specificity balance

F1-Score

Precision-recall
Harmonic mean

Sensitivity

Sensitivity
True Positive Rate

Specificity

Specificity
True Negative Rate

Dice/IoU

Segmentation accuracy
Region overlap measurement

Pneumonia Diagnosis Performance Comparison Example

Model	Accuracy	AUC	F1
X-ray Only	84.2%	0.87	0.82
Clinical Only	79.5%	0.83	0.78
X-ray + Clinical (Multimodal)	**91.3%**	**0.94**	**0.90**

Interpretability Challenges in Multimodal AI

Interpretation difficulties due to multimodal model complexity, utilizing attention visualization and SHAP

Need for Interpretability

Ensuring reliability of medical AI

- Supporting clinical decision-making
- Regulatory requirements (FDA, CE)
- Patient explanation responsibility
- Model debugging and improvement

Multimodal Interpretation Challenges

Sources of complexity

- Multi-layer fusion structure
- Inter-modal interactions
- Non-linear transformations
- High-dimensional representation space

Attention Visualization

Attention weight analysis

- Cross-attention maps
- Modal-specific contributions
- Heatmap overlay
- Temporal attention patterns

XAI Techniques

Explainable AI methodology

- SHAP: Shapley values
- LIME: Local approximation
- Grad-CAM: Gradient-based
- Integrated Gradients

Modal Importance Scoring

Feature Contribution Analysis

Contrastive Case Presentation

Case Study: Emergency Room Multimodal Triage

Integrating text (complaints), images (X-ray), and biosignals (ECG, BP) in the emergency room to determine triage priority

Patient Arrival
& Initial Assessment

Multimodal
Data Collection

Multimodal
AI Analysis

Risk Score
Calculation

Triage
Prioritization

Text

Chief Complaint: "Chest pain"
Duration: 2 hours
History: Hypertension

Imaging

Chest X-ray
Cardiomegaly findings
Suspected pulmonary edema

Biosignals

HR: 110 bpm
BP: 160/95
ECG: ST elevation

Integrated Analysis Results URGENT

- **Predicted Diagnosis:**High probability of acute myocardial infarction (87% confidence)
- **Risk Score:**9/10 (Urgent)
- **Recommended Action:**Immediate cardiology consultation, prepare for emergency cardiac catheterization
- **Processing Time:**3 seconds (real-time support)

Hands-on: Multimodal Model Implementation Practice

Implement image+text fusion model with PyTorch, data loader, and training loop

Step 1

Data Preparation

Load image and text paired dataset

```
dataset = MultimodalDataset(
  image_dir="./xrays",
  text_file="reports.csv"
)
```

Step 2

Define Encoders

Build separate encoders for image/text

```
img_encoder = ResNet50()
text_encoder = BERT()
```

Step 3

Fusion Layer

Feature concatenation and projection

```
fusion = nn.Linear(
  img_dim + text_dim,
  hidden_dim
)
```

Step 4

Training Loop

Loss function and optimization

```
loss = contrastive_loss(
  img_emb, text_emb
)
optimizer.step()
```

GitHub:
Multimodal-Medical

Colab Notebook:
Practice Examples

Dataset:
MIMIC-CXR

Future Directions: The Future of Multimodal Medical AI

New Modality Integration, Large-scale Pre-training, Real-time Clinical Application, Explainability Enhancement

Foundation Models

Large-scale Multimodal Pre-training

- Medical-specific GPT-4 level models
- Simultaneous learning of diverse modalities
- Few-shot adaptation capabilities
- General-purpose medical AI platform

New Modalities

Expanding Data Types

- Wearable sensor integration
- Microbiome data
- Social media health information
- Environmental factor data

Real-time Systems

Immediate Clinical Support

- Edge AI deployment
- Latency minimization
- Continual learning
- Mobile healthcare

Ethics and Regulation

Responsible AI Development

- Bias mitigation techniques
- Privacy-preserving learning
- International standardization
- Clinical validation protocols

Personalized

Preventive

Improved

Precision Medicine

Health Management

Healthcare Accessibility

Thank You!
Multimodal Medical AI: 다양한 데이터 통합으로 더 정확하고 포괄적인 의료 인공지능 구현

Ho-min Park
homin.park@ghent.ac.kr
powersimmani@gmail.com