Interpretability Challenges in Multimodal AI

멀티모달 모델의 복잡성으로 인한 해석 어려움, 어텐션 시각화 및 **SHAP** 활용

해석 가능성 필요성

의료 AI의 신뢰성 확보

- 임상의 의사결정 지원
- 규제 요구사항 (FDA, CE)
- 환자 설명 책임
- 모델 디버깅 및 개선

멀티모달 해석 어려움

복잡성의 원인

- 다층 융합 구조
- 모달 간 상호작용
- 비선형 변환
- 고차원 표현 공간

Attention 시각화

어텐션 가중치 분석

- Cross-attention 맵
- 모달별 기여도
- 히트맵 오버레이
- 시간적 어텐션 패턴

XAI 기법

설명 가능 AI 방법론

- SHAP: Shapley values
- LIME: 로컬 근사
- Grad-CAM: 그래디언트 기반
- Integrated Gradients

모달 중요도
점수

특징 기여도
분석

대조 사례
제시