서식 3 **공모 제안서[공모분야 #2 허위정보 대응 아이디어]**

*접수번호	

「허위정보 대응 아이디어」 공모전

제 안 서

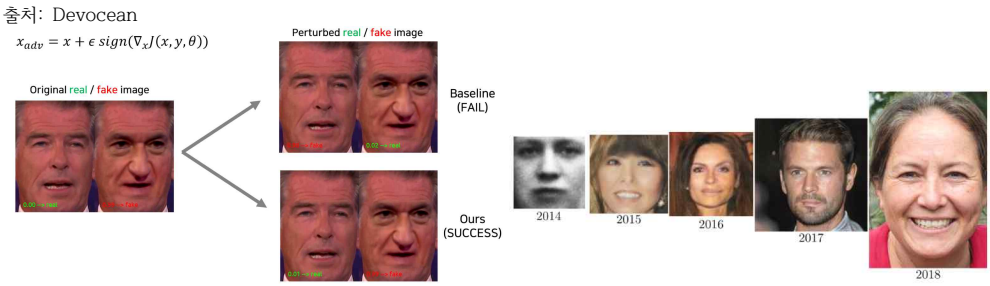
공모 제목	AI 멀티모달 딥페이크 검증 시스템
세부 분야	(1) 텍스트, 이미지, 영상 등으로 만들어진 허위 정보(가짜뉴스, 딥페이크 등)를 신속하게 탐지하는 방법에 대한 아이디어
제안 배경	

1. 딥페이크의 등장 및 사회적 심각성

딥러닝 기술의 발전으로 인해 딥페이크(Deepfake)와 같은 가짜 정보 생성 기술이 급속히 대중화되고 있다. 딥페이크는 주로 GAN(Generative Adversarial Networks) 같은 생성 모델을 활용해 특정 인물의 얼굴, 목소리, 제스처 등을 실감나게 조작하는 방식으로, 주로 영상이나 음성 콘텐츠에서 실현된다. 이러한 허위 정보는 기존의 가짜 뉴스보다 훨씬 정교하며, 사람들의 육안으로 진위를 가려내기 어렵다는 특징을 지닌다. 그 결과, 유명 인사의 가짜 발언이나 조작된 사건 영상을 통해 여론이 왜곡되는 경우가 있고, 특히나 2024년 교육부의 조사에 따르면 8.28~9.6까지 **딥페이크로 인한 청소년 성범죄 피해신고 건수는 238건에 달할정도로 성범죄에 악용되고 있으며, 특히 청소년을 대상으로 한다는 점에서 심각한 사회문제로 대두되고 있다.**

2. 딥페이크의 기술적 고도화로 인한 탐지의 어려움

출처: Devocean



딥페이크 기술이 발전하면서, 탐지 기술이 발전해도 **이를 우회하는 딥페이크 생성 방식이 빠르게 등장하고 있다.** 최신 딥페이크 생성 모델은 단순한 얼굴 합성 기술을 넘어, 음성 패턴, 감정 표현, 빛과 그림자 같은 세밀한 요소까지 실제처럼 재현하는 수준에 이르고 있다. 특히, 트랜스포머(Transformer)와 같은 최신 AI 아키텍처가 적용되면서, 단순한 탐지 알고리즘으로는 특정 영상 속 조작된 얼굴 움직임이나 음성 패턴을 판별하기가 더욱 어려워졌다. 결과적으로, 기존의 텍스트, 이미지, 영상 등의 **단일 모달리티 분석 방식만으로는 정교하게 합성된 딥페이크를 탐지하는 데 한계가 존재**하게 되었으며, 이로 인해 더 고도화된 탐지 시스템이 요구되고 있다.

3. 딥페이크 탐지를 위한 멀티모달 AI검증 시스템의 필요성

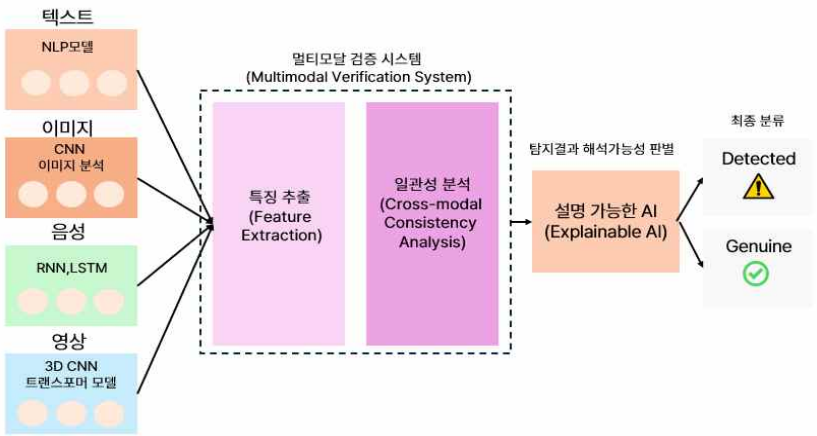
고도화된 딥페이크 탐지의 어려움을 극복하기 위해, 기존의 단일 모달리티 검증방식과 차별화된 **다양한 모달리티(텍스트, 이미지, 영상, 음성)를 동시에 분석하여 교차 검증하는 멀티모달 AI 검증 시스템**이 효과적인 방법이 될 수 있다. 멀티모달 시스템

은 딥페이크 생성 모델이 각기 다른 모달리티 간의 불일치를 자연스럽게 합성하기 어려운 점을 이용한다. 예를 들어, 영상 속 인물의 발언과 입술의 움직임이 미세하게 어긋나거나, 배경과 얼굴의 조명 불일치와 같은 특징을 찾아냄으로써 고도로 합성된 딥페이크도 효과적으로 탐지할 수 있다.

이러한 시스템은 CNN(Convolutional Neural Network)과 LSTM(Long Short-Term Memory), NLP(Natural Language Processing) 모델을 결합하여 텍스트-이미지, 음성-영상 간의 일관성을 분석하며, 딥페이크가 포함된 콘텐츠의 진위를 더욱 정확하게 판단할 수 있다.

아이디어 개요

<멀티모달 AI 검증 시스템 도식>



딥페이크 탐지를 위해 제안하는 멀티모달 AI 검증 시스템은 텍스트, 이미지, 음성, 영상 등 다양한 모달리티를 동시에 분석하여, **각 모달리티 간의 일관성을 교차 검증하는 방식**으로 딥페이크를 탐지하고자 한다.

멀티모달 검증 시스템은 딥페이크 생성 모델이 서로 다른 모달리티 간의 불일치를 자연스럽게 합성하기 어려운 점을 활용하여, 고도로 합성된 딥페이크 콘텐츠도 효과적으로 탐지할 수 있는 가능성을 제공한다.

예를 들어, 영상 속 인물의 발언과 입술 움직임이 미세하게 어긋나는지 분석하거나, 배경 조명과 얼굴 조명이 불일치하는지 확인함으로써, 단일 모달리티 탐지로는 식별하기 어려운 복합적인 허위 정보를 식별할 수 있다.

제안하는 시스템은 CNN(Convolutional Neural Network), LSTM(Long Short-Term Memory), NLP(Natural Language Processing) 모델을 결합하여, 텍스트-이미지, 음성-영상 간의 일관성을 분석하는 멀티모달 학습 방법론을 적용한다. 텍스트 분석은 뉴스 기사나 자막의 문맥과 주요 키워드를 추출하고, 이미지 분석은 얼굴 표정과 배경 요소를 인식하여 영상 콘텐츠와의 일관성을 평가한다. 또한, 음성 패턴과 입술 움직임을 비교하여 조작 가능성을 평가하며, 영상 내 시각적 요소 간의 불일치(예: 조명 불일치)를 검출함으로써 정밀한 탐지 기능을 수행한다.

<CNN이미지 분석 및 attetion 메커니즘의 이론적 고찰>

$$(X * W)(i,j) = \sum_m \sum_n X(i+m, j+n) \cdot W(m,n)$$

*이미지 특징 추출 및 크기 축소를 위한 2X2 커널 풀링연산

$$P(i,j) = \max(X(i+m, j+n))$$

셀프 어텐션 (Self-Attention) $\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

포지셔널 인코딩 (Positional Encoding) $PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$

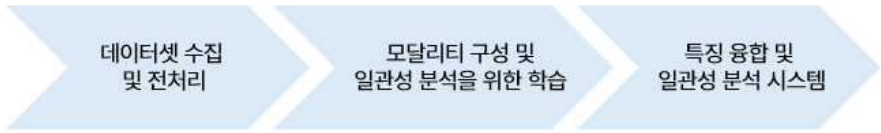
$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$

Q : 쿼리 벡터, K : 키 벡터, V : 값 벡터, d : 차원수, pos : 단어 위치, d : 임베딩 차원

세부 방안

<시스템 프레임워크>

프레임워크



(1) 데이터 수집 및 전처리

멀티모달 AI 검증 시스템은 딥페이크 영상자료를 판별하는데 가장 적절하고 목적적 합함으로 영상의 이미지, 영상, 음성 데이터셋 등을 수집한다.

모달리티	전처리 방식
텍스트	불필요한 기호, 불용어(stop words) 처리 및 토큰화와 BERT를 활용한 임베딩 처리
이미지	CNN 모델에 입력하기 위한 이미지 크기 조정, 노이즈 제거, 데이터 증강(augmentation)
음성	주파수 스펙트럼 분석, Mel-spectrogram 변환 및 입력
영상	프레임 단위 분할 및 얼굴 특징 추출 및 표정분석

(2) 멀티모달 AI 모델 구성

각 모달리티의 AI 모델을 구성하고 모달리티 간 일관성 분석을 위한 학습

모달리티	모델 학습
텍스트	BERT(Bidirectional Encoder Representations from Transformers): 텍스트 데이터에서 문맥 정보를 추출하고 허위 정보를 탐지하기 위한 모델로 사용. 문장의 일관성, 키워드, 문맥적 유사성 등을 평가. 기술적 구성: BERT 모델의 트랜스포머 인코더를 사용하여 텍스트 임베딩을 추출하며, 이를 후속 멀티모달 검증 시스템에 전달.
이미지	ResNet 또는 Inception 모델: CNN 기반 모델을 사용하여 이미지에서 얼굴 표정, 조명 조건, 배경 정보를 추출하고 합성 여부를 판단. 기술적 구성: 합성된 딥페이크 이미지의 특징을 학습하고, 인물의 얼굴 특징 및 배경 정보가 영상이나 텍스트와 일치하는지를 확인.
음성	RNN 또는 LSTM 기반의 음성-텍스트 동기화 모델: 음성과 영상의 입술 움직임이 일치하는지를 분석하여 딥페이크 여부를 판별. 기술적 구성: ASR(Automatic Speech Recognition) 모델을 통해 음성을 텍스트로 변환하고, 텍스트-음성 동기화를 평가하여 탐지 모델에 전달.
영상	3D CNN 또는 Vision Transformer: 영상의 프레임 간의 시간적 연속성을 분석하여 인물의 얼굴 표정 변화와 배경 조명이 일관성 있게 나타나는지를 분석. 기술적 구성: 영상 프레임에서 표정과 입술의 움직임을 추적하여 음성과 일치하는지, 텍스트와 일관된 행동이 나타나는지를 평가.

(3) 특징 융합 및 일관성 분석

어텐션 메커니즘(attention mechanism)을 활용하여 각 모달리티의 중요도를 동적으로 조정한다. 또한 특정 콘텐츠에서 어떤 모달리티가 더 중요한 정보를 포함하는지 학습하고, 그에 따라 가중치를 부여한다.

텍스트 모달리티: $f_{text} \in \mathbb{R}^{d_t}$ (BERT 기반 텍스트 임베딩)

이미지 모달리티: $f_{image} \in \mathbb{R}^{d_i}$ (CNN 기반 이미지 특징)

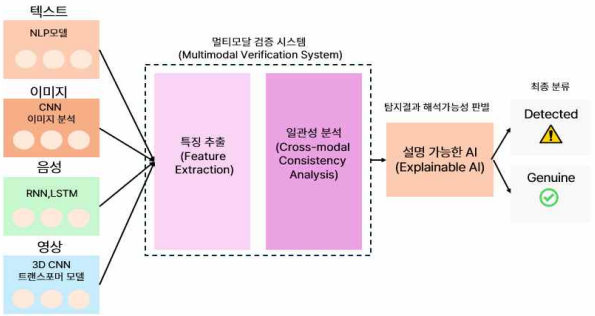
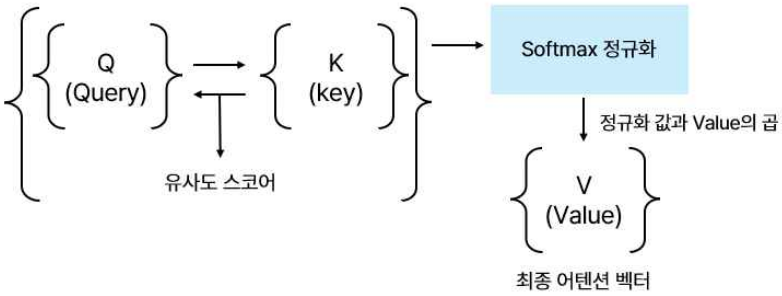
음성 모달리티: $f_{audio} \in \mathbb{R}^{d_a}$ (LSTM 기반 음성 임베딩)

영상 모달리티: $f_{video} \in \mathbb{R}^{d_v}$ (3D CNN 또는 비디오 트랜스포머 특징)

이 각각의 벡터들을 공통 임베딩 공간으로 매핑하기 위해 아래와 같이 선형 변환을 적용한다.

$$f'_{modality} = W_{modality} \cdot f_{modality} + b_{modality}$$

	각 모달리티의 특징 벡터 간 상관관계를 학습하기 위해 셀프 어텐션(self-attention)으로 상대적 중요도를 생성한다. $$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$ $$F_{fusion} = \sum_{modality} \alpha_{modality} \cdot f'_{modality}$$ 어텐션 메커니즘을 통해 생성한 중요도에 따라 통합된 특징 벡터 F를 생성한다. **(4) 불일치 탐지 알고리즘** **(4)-1 유사도 계산 및 가중평균** 모달리티 간 불일치 여부를 판별하기 위해 벡터 간 유사도를 코사인 유사도(Cosine Similarity)를 통해 측정 후 가중평균하여 전체 일관성 점수 S를 계산한다. $$\text{Similarity}(f_a, f_b) = \frac{f_a \cdot f_b}{\|f_a\|\|f_b\|} \qquad S = \sum_{(a,b)\in\text{pairs}} w_{(a,b)} \cdot \text{Similarity}(f_a, f_b)$$ **(4)-2 불일치 최종 판별** 일관성 점수 S가 임계값보다 낮을 경우 허위정보로 분류하며, 신뢰도 점수를 산출한다. $$\text{Label} = \begin{cases} \text{Fake}, & \text{if } S < \tau \\ \text{Real}, & \text{otherwise} \end{cases} \qquad \text{Confidence} = \frac{1}{N}\sum_{modality} \alpha_{modality} \cdot \text{Similarity}(f_{modality}, F_{fusion})$$ ***임계값은 딥페이크를 판별하는 가장 중요한 설정값으로, 신뢰도 평가 이외에도 추후 각종 지도학습 평가지표(Accuracy, Precision, F1-score) 등을 활용하여 최적화된 값을 설정할 필요가 있다.**
장점 및 기대효과	**(1) 영상매체 탐지 최적화** 기존의 단일 모달리티 기반 탐지 시스템은 특정 데이터 유형에만 최적화 되어있어, 텍스트나 이미지 같은 단일 요소에서만 딥페이크를 탐지하는 한정적인 성능을 보였다. 하지만 실제 탐지하기 어려운 딥페이크는 동영상이 많으며, 멀티모달 AI 검증시스템은 이를 탐지하기 위해서는 영상 속 텍스트, 음성, 이미지 간의 상호작용과 일관성을 중심으로 평가하여 미묘한 분석까지 가능하다는 점에서 장점이 있다. **(2)다양한 연구를 통합할 수 있는 적응성과 현실성** 현재 제시한 시스템은 텍스트, 이미지, 음성, 영상 뿐이지만, 현재 딥페이크 탐지를 위해 연구가 이뤄지고 있는 생체신호, 감정분석과 같은 새로운 기술을 통합하기에 현실적으로 용이하다. 데이터와 요구사항에 따라 모달리티의 가중치를 조정하거나, 실제 성능에 따라 어떤 모달리티에 더욱 높은 판별 능력이 있는지를 분석하여, 적응성을 확보할 수 있다. 일례로, KAIST에서 딥페이크 영상을 잡아내는 '카이캐치(KaiCatch)' 소프트웨어는 얼굴의 미세 변형과 코, 입, 얼굴 윤곽 등 생체 기하학적 왜곡을 탐지하는 모바일 앱인데, 이와 같은 탐지 기술 또한 멀티모달 AI 검증 시스템의 최적화를 위해 기술을 탑재할 수 있다. 2024년 9월 방송통신위원회의 딥페이크 성범죄 영상물 대응 전문가 토론회에서는 **딥페이크 성범죄 영상물을 탐지하는 기술에 국가지원이 필수적이라는 점을 언급한 바 있으며, 이에 대한 정부의 기술 연구지원 검토가 이뤄지는 것을 밝힌 바 있다.** **-> 멀티모달 AI 검증 시스템에서 활용되는 어텐션 메커니즘은 이미 컴퓨터 비전, 음성처리,번역 등에서 활발하게 활용되고 있으며, 딥페이크 문제 해결을 위한 다양한 모달리티 분석에 어텐션 메커니즘을 적용시키는 것은 충분히 실현가능한 원리다.**

	사회적 파급효과 **(1) 개인 프라이버시 보호** 딥페이크로 인한 개인 정보 유출 및 명예훼손의 문제를 줄일 수 있으며, 피해자가 허위 콘텐츠로 인한 정신적, 경제적 피해를 받지 않도록 지원할 수 있다. **(2) 딥페이크 예방을 위한 국제적인 기술적 협력 촉진** 멀티모달 탐지 방식은 다방면의 전문가가 공동으로 연구해야 할 필요가 있기 때문에 국제사회 각 분야의 전문가가 공동으로 허위 정보 탐지 및 방지 기술을 개발하고 공유할수 있다. 딥페이크 문제는 우리나라에만 국한된 것이 아닌, 전세계적으로 악영향을 주고 있는 심각한 문제이기 때문에 국제적 기술적 협력을 촉진할 수 있다. **(3) 지속적 연구개발의 기반 마련** 딥페이크 생성 기술은 갈수록 정교해지는만큼 이제는 단일화된 모달리티로 탐지하는 것보다 멀티모달의 체제가 들어서야한다. **모든 딥페이크 연구는 '더욱 정교한 딥페이크 탐지'라는 공통의 목표를 두고 있으므로, 각 분야에서의 연구를 통합하는 시스템이 도입되어야 지속가능하게 성능을 높일 수 있는 통합 연구가 진행될 수 있다.**
개념도	**<멀티모달 검증 시스템의 도식화>**  **<어텐션 메커니즘의 도식화>** 
기타 사항	참고 문헌 1. DeepFake Detection by Analyzing Convolutional Traces 2020 - Luca Guarnera 2. 딥러닝기반의 눈깜빡임 추적 및 바이오 리듬 분석을 통한 가짜 동영상 탐지 정택현,김기천*건국대학교 3. Multimodal Deep Learning - Jiuquan Nigiam 4. A review on the attention mechanism of deep learning - Zhaoyang Niu