

18조 이세연 이수연 강성은 김건우 문승기

DEEPTECTOR

AI기반 실시간 딥페이크 탐지 플랫폼

Overview

‘DEEP’TECTOR

[Deepfake + Detector (protector)]

딥페이크 이미지를 실시간으로 탐지하고 차단하여
안전한 디지털 환경을 구축하는 AI 플랫폼!

PLAY DEMO



Background

최근 N번방 사건과 같은 디지털 성범죄가 크게 증가하고 있다
그 중 한축은 딥페이크로, 피해자와 가해자의 상당수가 청소년이다

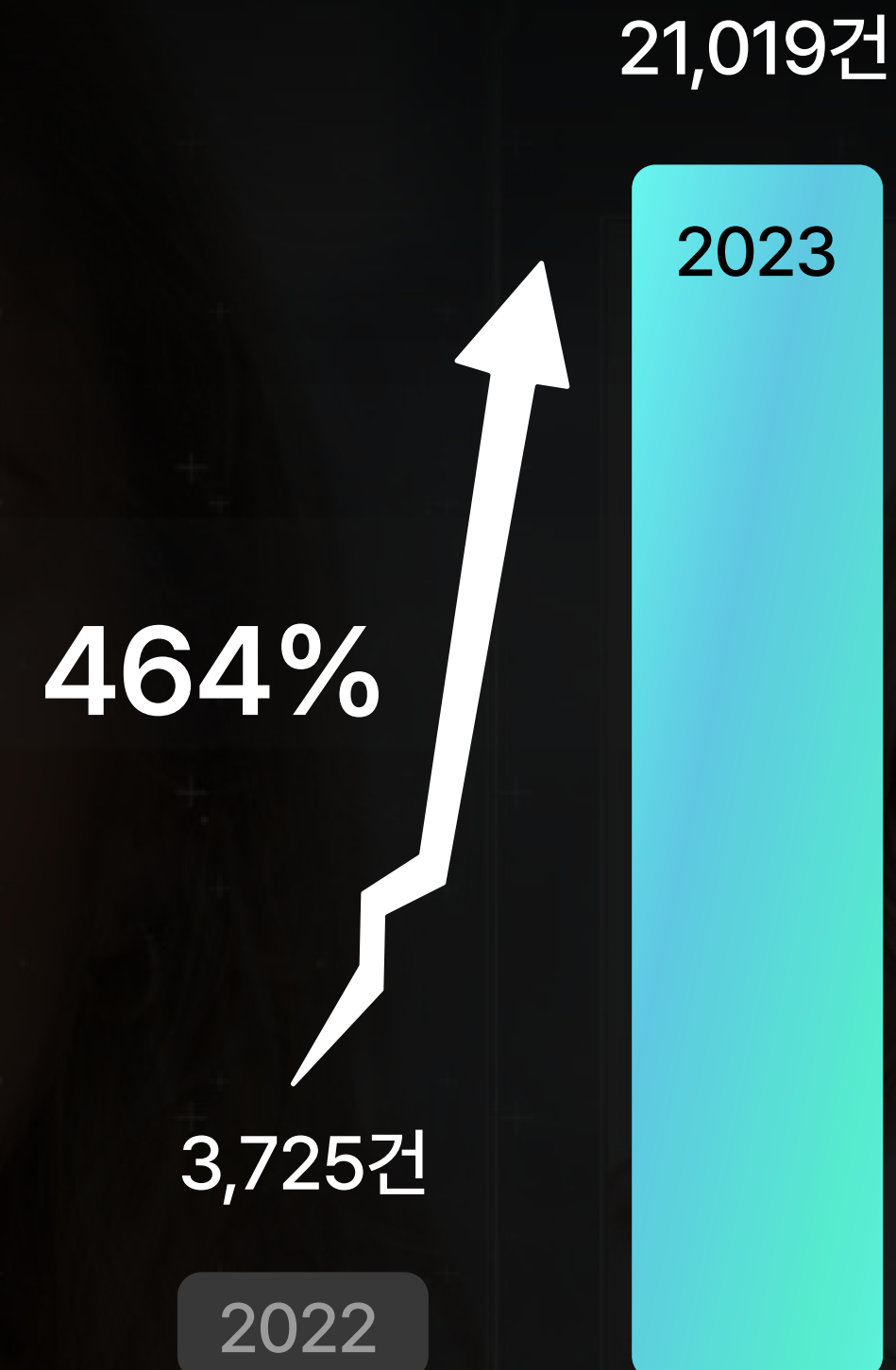
"30초만에 딥페이크 영상 완성"...더 교묘해진 'N번방' 범죄

2024년 05월 24일 오후 5:16

[앵커]

지난 2020년 우리 사회를 충격에 빠뜨렸던 'N번방 사건' 이후 각종 대책이 마련됐지만, 디지털 성범죄는 더 발 빠르게 진화하고 있습니다.

특히 최근 벌어진 '서울대판 N번방'은 여성 얼굴 사진에 음란물을 합성해 유포했는데, **최근 이 같은 '딥페이크' 기술을 이용한 성범죄가 급증하는 추세입니다.**



Q. 딥페이크 영상물 건수

Background

디지털 성범죄로 인해 삭제 요청을 해도 불법 콘텐츠의 확산과 2차 피해를 막기 어렵다

[단독] 삭제 요청 불응한 불법촬영물 16만건...“100% 차단은 불가능”

공성운 기자 (niceball@sisajournal.com) | 승인 2024.08.12 15:32

최근 3년 간 삭제지원 62만건 중 20~30% 불응...방심위 차단해도 우회접속 가능성 존재

정부가 최근 3년 간 삭제 요청한 불법촬영물 62만여 건 중 약 16만 건이 요청을 묵살한 것으로 나타났다.

불법촬영물은 지금도 유포되고 있을 가능성을 배제할 수 없는 상황이다.

9일 여성가족부 디지털성범죄방지와 관계자는 시사저널에 “2021년부터 작년까지 매년 삭제 지원한 불법촬영물 중 20~30%의 경우 (삭제 요청에) 불응하고 있다”고 했다. 디지털성범죄를 전담하는 여성가족부 산하 디지털성범죄피해자지원센터(디성센터)는 자체 검색시스템을 통해 온라인상에 퍼진 불법촬영물을 찾아 사이트 측에 삭제 요청을 하고 있다. 그래도 완전한 제거가 힘든 상황이다.

'성착취물 삭제' 구글에 요청했더니...답변만 1년 걸렸다

전혼잎 기자 구독 + 입력 2022.12.08 09:00

신고를 했다고 바로 삭제가 되지도 않았다. 그가 구글에 답변받기까지 기다린 시간은 무려 1년. 김씨는 “피해자들에게 구글은 거대한 유포 웹사이트에 불과하다. 그런 면에서 구글은 최악

“한국의 온라인 성폭력 생존자들이 구글의 느리고 복잡한 콘텐츠 삭제 요청 시스템으로 인해 더욱 큰 고통을 겪고 있다”

국제앰네스티는 국내 온라인 성폭력 피해 생존자와 활동가를 대상으로 설문조사를 실시한 결과, “한국의 온라인 성폭력 생존자들이 구글의 느리고 복잡한 콘텐츠 삭제 요청 시스템으로 인해 더욱 큰 고통을 겪고 있다”라고 8일 밝혔다. 세계 최대 검색엔진인데도 비동의 성적촬영물 신고 절차를 찾기가 어려운 데다, 신속히 처리도 되지 않아 성착취 영상이 온라인에서 확산하고 있다는 것이다.

Background

딥페이크 탐지 솔루션에서 개인의 접근성을 강화하고 불법 콘텐츠를 신속하게 사전에 예방할 수 있는 플랫폼이 필요하다



Deep Brain

딥페이크 탐지 솔루션으로
사용자가 사이트에 직접 사진을 업로드해야
탐지가 가능하며 **기관(수사기관, 기업 등)**
중심으로 운영되어 개인 피해자가
접근하기 어려운 문제점이 있다.



삭제 지원 시스템

디지털 성범죄 피해자 지원센터의
수동 처리로 인해 시간이 오래 걸리거나
영구적으로 삭제되지 않는 문제가 있다.
또한, 피해자의 신상 정보는 삭제가 불가능해
2차 피해가 발생할 우려가 있다.



카카오톡

카카오톡 오픈채팅의 이미지 검사 기
능은 채팅 환경에서 **검사 시간 소요가**
상당히 길어, 실시간 소통이 중요한 채
팅 환경에서 **사용성이 떨어진다.**



DEEPTECTOR

AI기반 실시간 딥페이크 탐지 플랫폼

ASIS

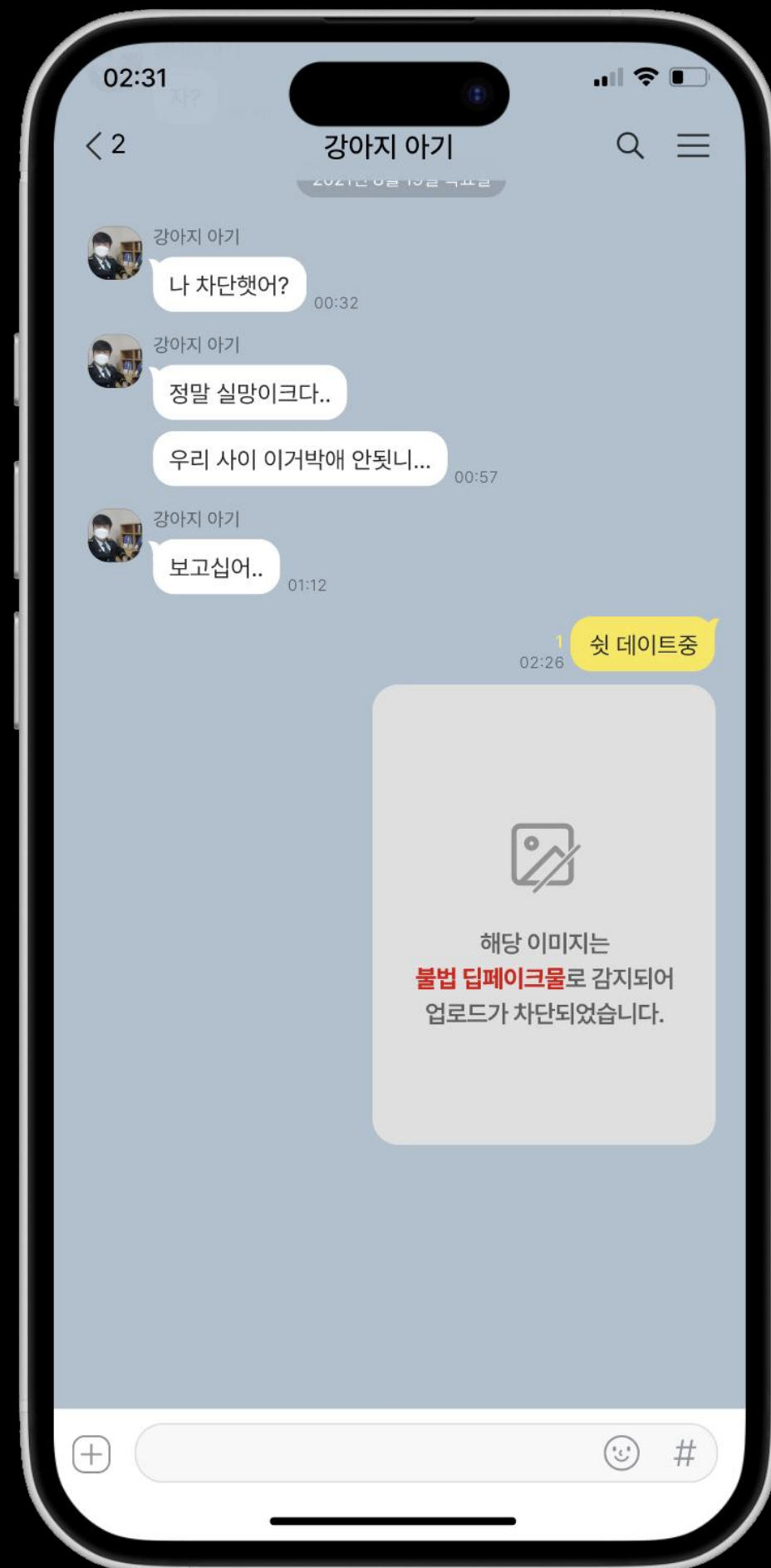
기존의 딥페이크 탐지 기술은 사후 처리 중심의
수동적 접근으로 인해 개인이 사용하기 어렵고
활용성 또한 제한적이다. 예방보다는
사후 처리에 초점이 맞춰져 있어 범죄 예방에는
효과가 미흡한 실정이다.

TOBE

이미지 업로드 전 **AI를 활용해 자동으로** 딥페이크를
사전 차단하고, 안전한 디지털 소통 환경을 제공한다.
이미지가 **서버에 저장되기 전** DEEPTECTOR API가
검사를 수행해 검증된 이미지만 서버에 업로드 되도록 한다.

Key-Function

인공지능 VIT 기술을 활용한 딥페이크 탐지 플랫폼



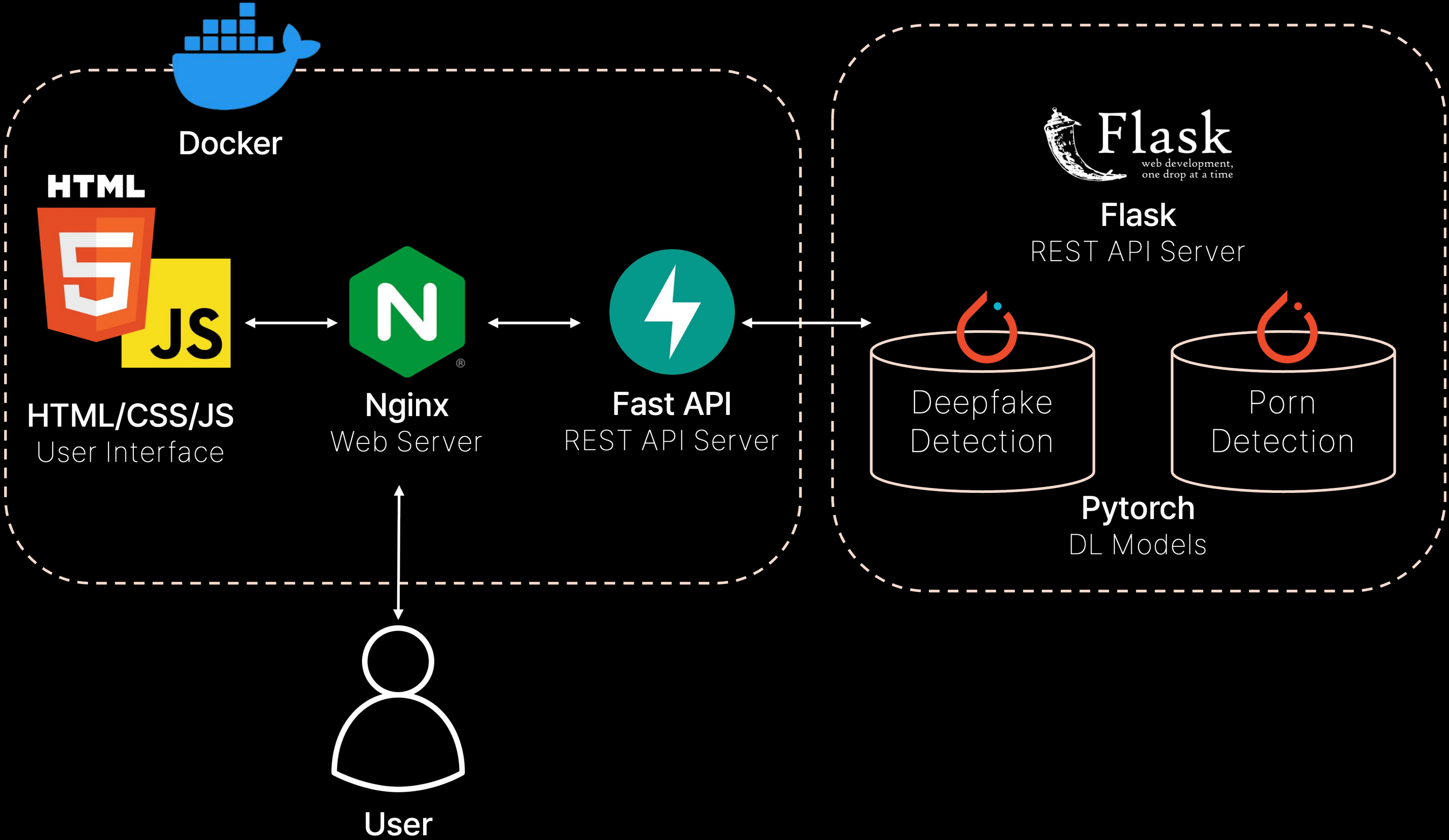
for Business

- 딥페이크 **콘텐츠 업로드 전에 실시간 탐지**하여 유포를 사전에 방지한다.
- 기존의 사후 신고-삭제 방식 대신, **사용자 요청 없이 자동 필터링**이 이루어진다.

for Consumer

- 딥페이크 노출이 걱정되는 사용자가 자신의 사진을 등록해, **해당 이미지와 유사한 딥페이크 사진이 있는지 검색**할 수 있다.

System Architecture



Key-Tech

1 딥페이크 이미지의 특징 추출

- Kaggle deepfake and real images 데이터셋에서 추출한 이미지 아티팩트(Image Artifact)



얼굴과 몸의 미세한 색감 차이



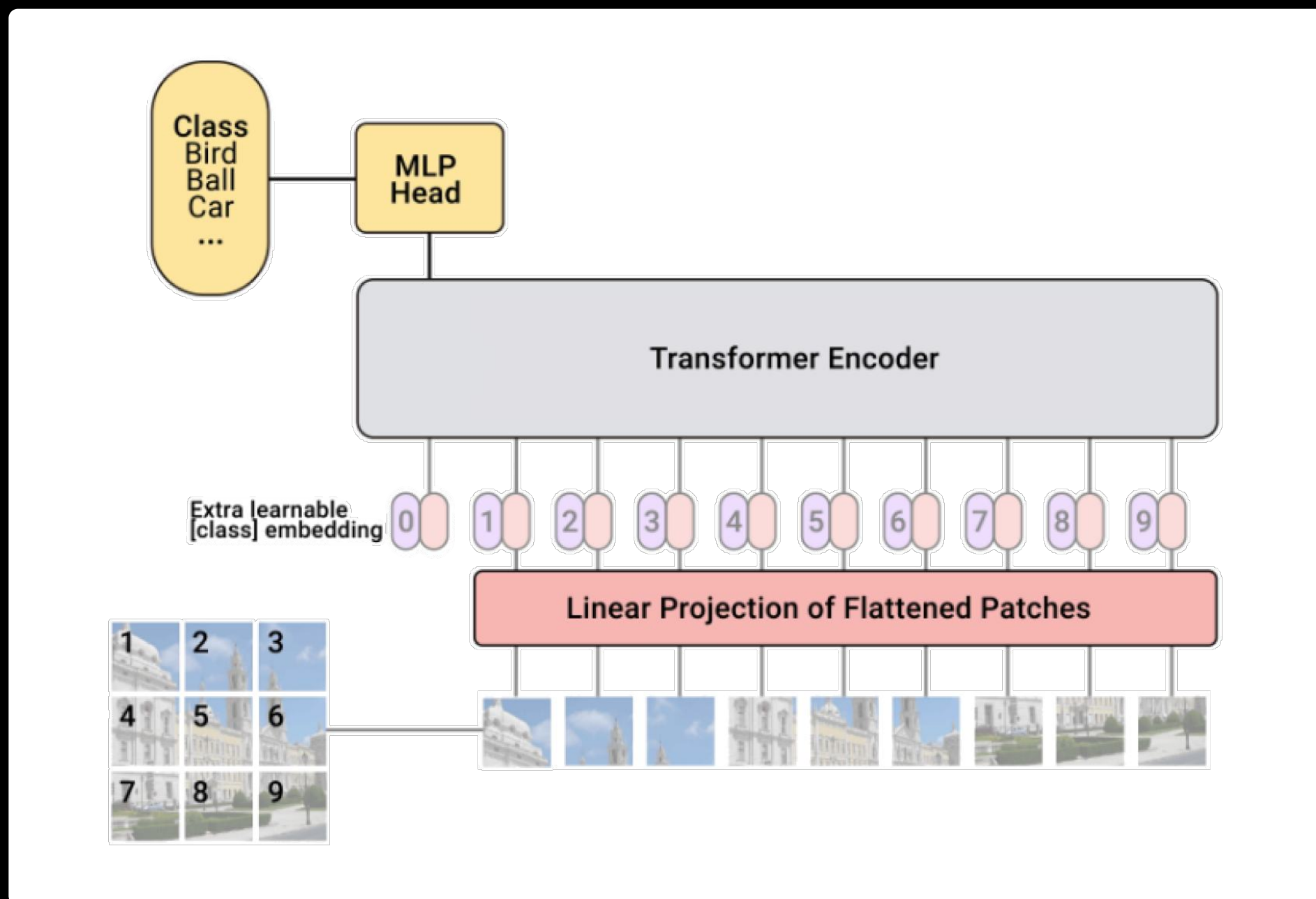
시각적 합성 오류



→ 이미지 아티팩트(Image Artifact)를 잘 감지할 수 있는 모델을 구축

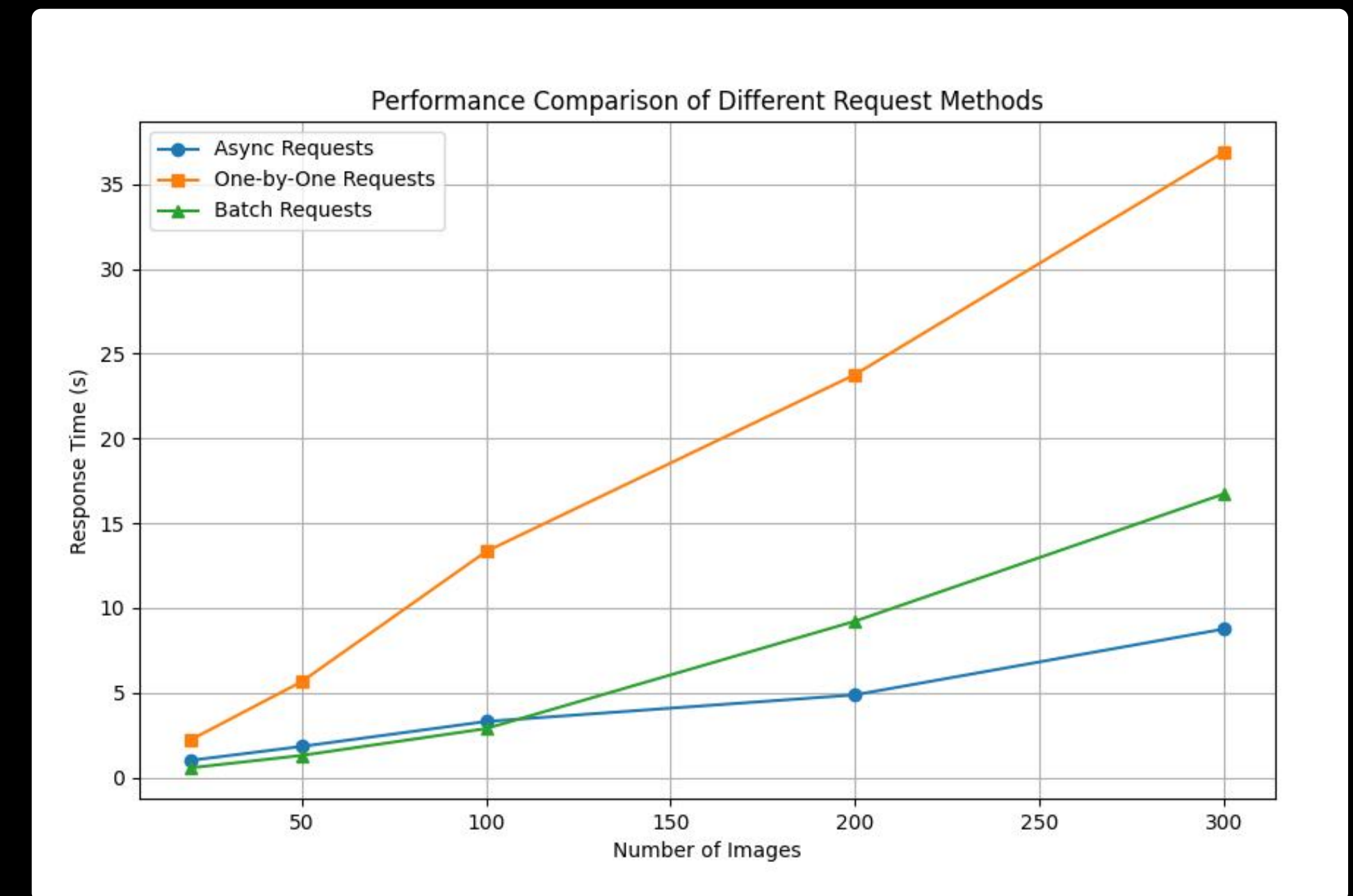
2 딥페이크 감지 Vision Transformer 모델

- Transformer-Based Model의 Vision Transformer(Vit) 사용
- 복잡한 이미지에서 특징추출에 강함, 이미지의 전체적인 특징 반영 가능
- 딥페이크 탐지 프로토타입 구현에 적합한 인공지능 모델으로 판단, Vit모델 사용



3 DL Model Serving Technique

- 모델 구축 후 API서버 구축 및 서빙
- 배치(Batch) , 비동기 방식을 이용해 여러 이미지를 한번에 처리
- 배치로 50장의 이미지 1.35초 안에 처리 가능



Sub-Function

사용자가 직접 신고 대상을 찾는 불편함을 줄이고,
자동으로 해당 플랫폼에 신고가 전달되어 빠르고 편리하게 조치가 이루어지도록 한다




개인 검색 기능

- 사용자가 자신의 사진을 업로드하면, 유사한 얼굴이 포함된 이미지와 영상을 **검색하여 개인화된** 결과를 제공한다.
- 탐지 결과를 "당신의 얼굴과 유사한 영상이 n건 발견되었습니다. 확인하시겠습니까?"와 같은 메시지로 사용자에게 알린다.

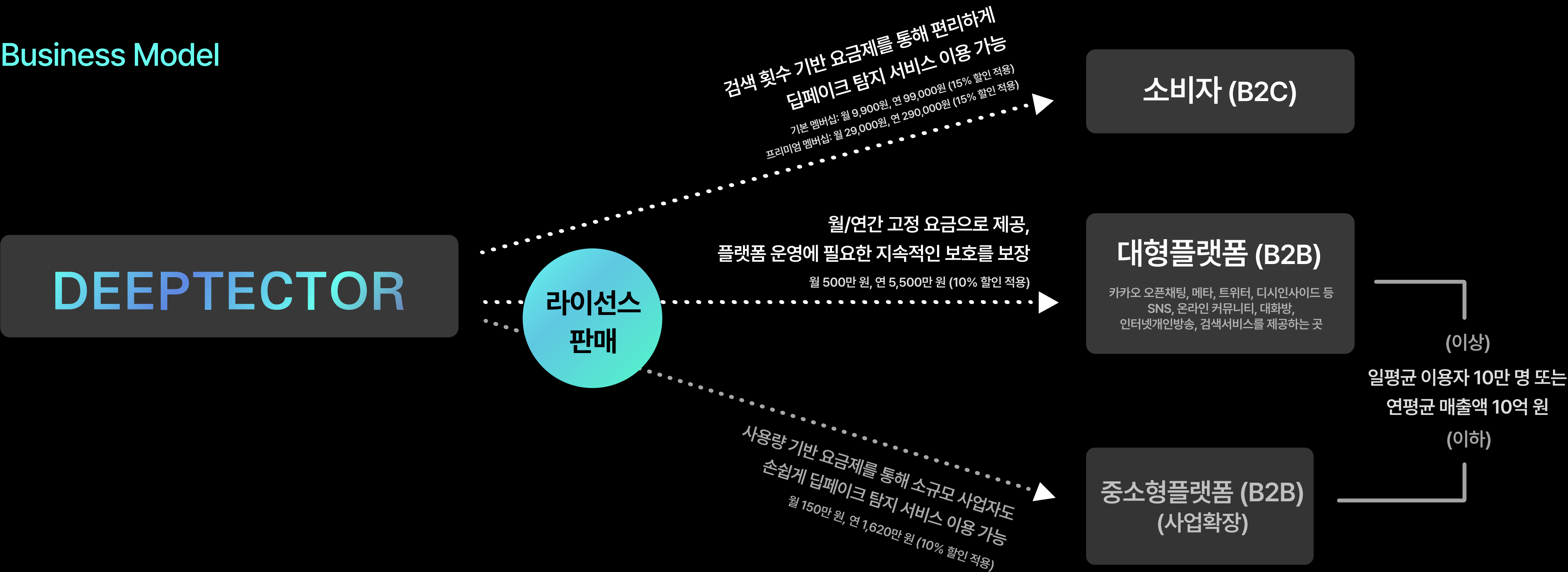
자동 알림 기능

- 페이크 영상 감지 시, 해당 플랫폼과 연동하여 사용자에게 즉시 메시지를 전송해 **빠른 대응**을 돕는다.

Benefits

 사용자 보호 강화 불법 딥페이크 및 음란물 전송을 사전 차단해 안전한 대화 환경 제공	 법적 문제 예방 법적 책임 경고를 통해 사용자가 불법 콘텐츠 전송을 자제하도록 유도	 신뢰성 있는 플랫폼 구축 플랫폼의 안전성과 사용자 신뢰도 향상
---	--	--

Business Model



DEEPTECTOR

Q&A

AI기반 실시간 딥페이크 탐지 플랫폼