2024.12.27

1

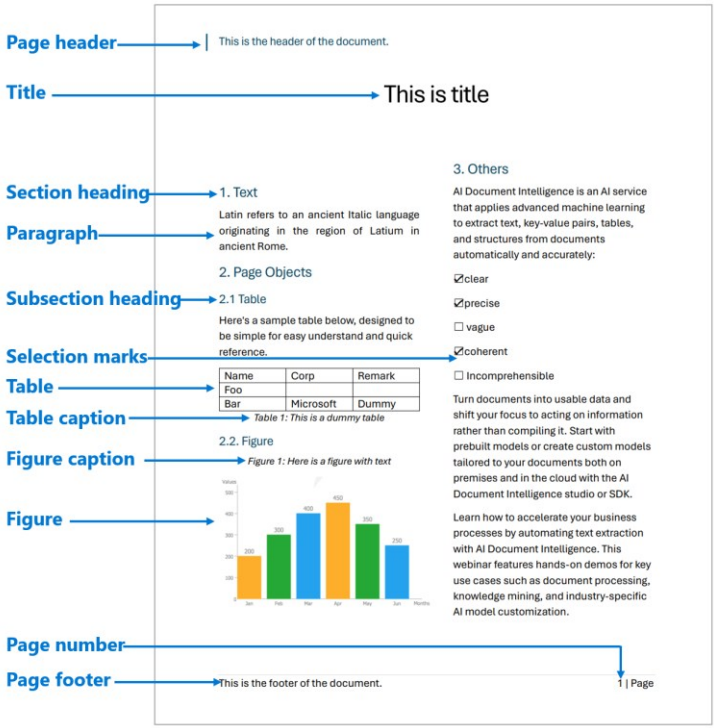
Document Layout Analysis

국립한국해양대학교 인공지능응용연구실 여지민

E-mail : jiroungree@naver.com

목차

Document Layout



- 문서에는 여러 섹션(ex: 본문, 제목, 표, 이미지, 주석)이 존재
- OCR(Optical Character Recognition)만으로 처리할 경우, 이 구조가 무시되어 텍스트가 뒤섞이는 문제가 발생

DLA(Document Layout Analysis) 기술이란?

- 문서의 구조와 레이아웃을 이해하기 위해 콘텐츠가 어떻게 공간적으로 배치되어 있는지를 분석하는 기술 및 과정

- 문서 내의 텍스트, 이미지, 테이블, 그 외 다른 요소의 위치를 구별하고, 제목, 장, 절 등 문서의 전체적인 구조를 구분

전통적인 기법(Rule-based)

VS

ML기반 기법(Machine Learning-based)

전통적인 기법(Rule-based)

1. Connected Component Analysis (CCA)

- 이진화된 이미지를 픽셀 단위로 분할한 뒤 서로 **연결된 영역을 감지**하여 글자, 단어, 문단 등의 레이아웃을 찾는 데 사용

- 글자, 기호 등 세부 구성 요소를 정확히 구분 가능

- 노이즈, 비정형 레이아웃에서는 오류가 발생

전통적인 기법(Rule-based)

2. Projection Profile

- 이미지의 수평 또는 수직 방향에서 **픽셀 밀도를 계산**해 텍스트 영역을 구분
- 텍스트 정렬 감지와 단순 레이아웃 문서에 적합
- 비정형 레이아웃에서는 한계가 발생

전통적인 기법(Rule-based)

3. Run Length Smoothing Algorithm (RLSA)

- 픽셀 간 간격이 특정 임계값 이내라면 연결하여 연속적인 텍스트 영역으로 묶어서 구분
- 비교적 간단한 연산과 문단 단위로 텍스트 영역을 묶는 것에 효과적
- 픽셀 간 간격 임계값 설정에 민감하며, 개별 구성 요소 구분 어려움

ML기반 기법(Machine Learning-based)

1. Convolutional Neural Networks (CNNs)

- **이미지 특징**을 학습하여 텍스트와 비텍스트 영역을 감지하여 구분
- 표, 이미지, 텍스트 등의 복잡한 레이아웃을 처리 가능
- 데이터의 크기와 연산량이 비교적 많음

ML기반 기법(Machine Learning-based)

2. Transformer 기반 모델

- 문서 이미지의 **텍스트**와 **시각적 정보**를 함께 처리

- 텍스트와 이미지 요소 간의 관계를 이해하여 정확하고 유연함

ML기반 기법(Machine Learning-based)

3. Recurrent Neural Networks (RNNs)

- 문서의 **순차적 요소**(텍스트 흐름)를 분석하여 페이지의 **논리적 구조**를 이해
- 텍스트, 이미지, 표 등 다양한 레이아웃 요소들의 **상호작용**을 분석
- 텍스트가 길어질수록 정확도가 떨어지고 연산처리가 많아짐

주요 DLA 모델

1. LayoutLM
2. DiT (Document Image Transformer)
3. Mask R-CNN
4. DocLayout-YOLO

LayoutLMv3

- 문서 이미지의 텍스트와 시각적 요소를 동시에 이해하기 위해 설계된 Transformer 기반 모델
- 텍스트 데이터 뿐 아니라 **텍스트의 위치**(Bounding Box 좌표)와 **이미지 정보**를 함께 입력으로 사용하는 멀티 모달 학습
- 잡지, 학술 논문과 같은 구조가 복잡한 문서에서 뛰어난 성능
- Google Document AI에 활용

DiT (Document Image Transformer)

- 문서 이미지의 구조와 내용을 이해하기 위해 설계된 **ViT(Vision Transfomers)** 기반 모델
- 문서 이미지의 **자기 감독 학습(Self-supervised Learning)**으로 학습
- LayoutLM과 달리 텍스트 OCR 없이 **이미지 정보만으로 학습**이 가능

Mask R-CNN

- Object detection과 Segmentation에 모두 활용 가능한 모델
- DLA에서 문서의 텍스트, 표, 이미지 영역을 감지하고 분리하는 데 사용
- 잡지, 보고서와 같은 문서에서 **텍스트와 비텍스트 구분**에 높은 성능

DocLayout-YOLO

- **YOLO(You Only Look Once)** 객체 탐지 알고리즘을 기반으로 설계된 DLA 특화 모델
- YOLO의 속도와 정확성을 활용하여 문서 레이아웃 분석을 효율적으로 수행
- 높은 FPS로 **실시간 분석**이 가능하고, 다양한 문서 유형에서 **일반화 성능**이 우수

한글 문서에 대한 DLA 모델 설계

- 한글 문서에 대한 DLA 모델을 구축하려면, OCR, 레이아웃 요소 탐지, 딥러닝 기반의 문서 구조 분석 기술을 결합
- 한글에 특화된 형태소 분석 및 OCR 기술을 활용
- 조건에 부합하는 알고리즘 선택
- 모델 학습을 위한 한글 문서 데이터셋 설정 및 데이터셋 전처리 과정 필요

2024 한국소프트웨어종합학술대회 포스트 논문 보고

1. 자연어 기반 이상상황 탐지를 위한 LLM 적용 의사결정트리 알고리즘 - 한국전자기술연구

2. 비전 언어 모델을 활용한 복합 얼굴 표정인식 - 한양대학교

자연어 기반 이상상황 탐지를 위한 **LLM** 적용 의사결정트리 알고리즘 - 한국전자기술연구원

- **Vision-Language Model(VLM)**을 통해 영상 데이터를 자연어로 설명하고, **Large Language Model(LLM)**과 의사결정 트리를 활용하여 이상상황을 인지하는 알고리즘을 제안

Vision 인식 기반 이상상황 탐지

영상 데이터 수집 → 사전 정의된 이상상황 학습 → 실시간 영상 분석 → 사전 정의된 상황 매칭 여부 판단 → 이상상황 알림

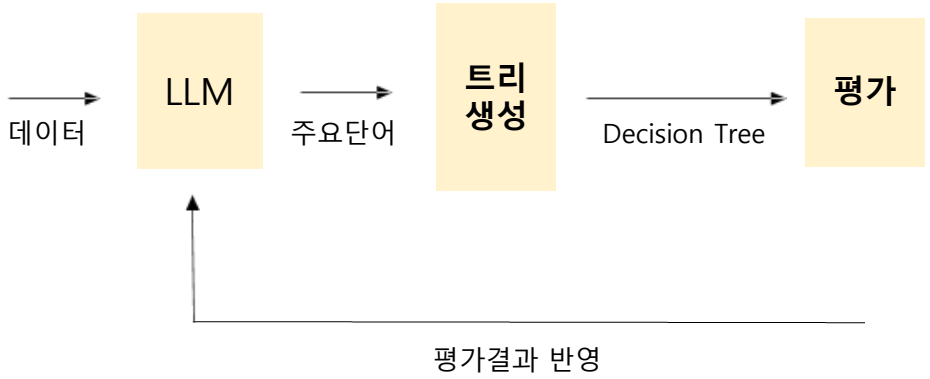
LLM 기반 자연어 활용 이상상황 탐지

영상 데이터 수집 → **VLM**을 활용해 영상 데이터를 자연어로 변환 → **LLM**으로 주요 키워드 추출 → 의사결정트리를 생성하여 위험 여부 판단 → 이상상황 알림

학습 단계

출입 금지 구역으로 접근 중인 사람. - 위험 레벨: 3
사람이 외딴 길목에서 다른 사람의 가방을 빼앗는다. - 위험 레벨: 4
도심 한복판에서 누군가 쓰러져 있다. - 위험 레벨: 5

<데이터: 전문가 지식>

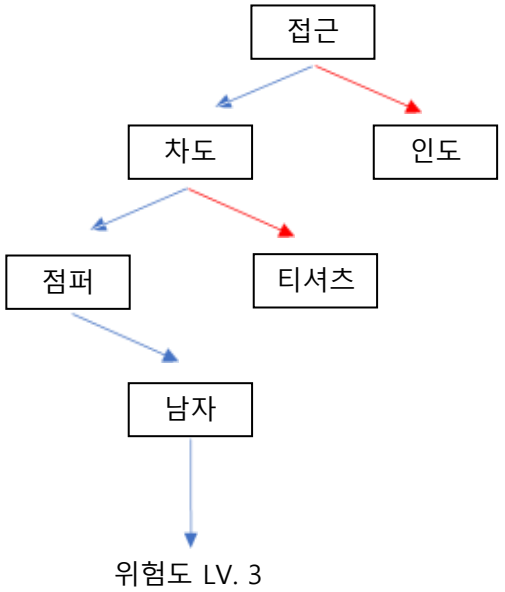


추론 단계

영상 스트림

VLM

11:13분 13번CCTV
노란색 점퍼를 입은 남자가
차도를 걷고있다.



특정단어가 문장에
포함되었는지를
기반으로
자연어 TREE 구축

결론

- LLM이 추출한 중요한 단어들이 모델 성능을 효과적으로 개선함을 입증

- 상황 탐지의 한계를 보완할 수 있는 가능성을 제시

비전 언어 모델을 활용한 복합 얼굴 표정인식 - 한양대학교

- 얼굴 이미지에서 단일 감정을 분류한 후, 비전-언어 모델로 다양한 감정 라벨을 생성해 복합 감정 라벨을 도출하는 알고리즘을 제안

서론

- 얼굴 감정 인식(Facial Emotion Recognition)은 인간의 감정을 얼굴을 통해 이해하는 기술로 전통적으로는 Happy, Sad, Anger 등과 같은 단일 감정 인식에 중점을 두고 발전

- 하지만 실제 감정은 단순하지 않으며, 여러 감정이 동시에 표현되는 복합적인 상태를 포함하는 경우가 많음

제안 방법

- VSM 모델을 활용하여 복합적인 얼굴 감정 라벨을 예측하는 체계를 구축

- 단일감정의 한계를 넘어 복합적인 감정상태를 더 정교하게 예측

감정 표현 생성 모듈

- 최신 VSM 모델인 LLaVA-NeXT를 사용하여 입력된 얼굴 이미지로부터 10개의 감정 표현을 생성

단일 감정 인식 모듈

- POSTER-V2 모델을 사용하여 7가지 단일 감정(Happy, Sad, Anger, Surprise, Fear, Disgust, Neutral)을 분류

감정 표현 임베딩 모듈

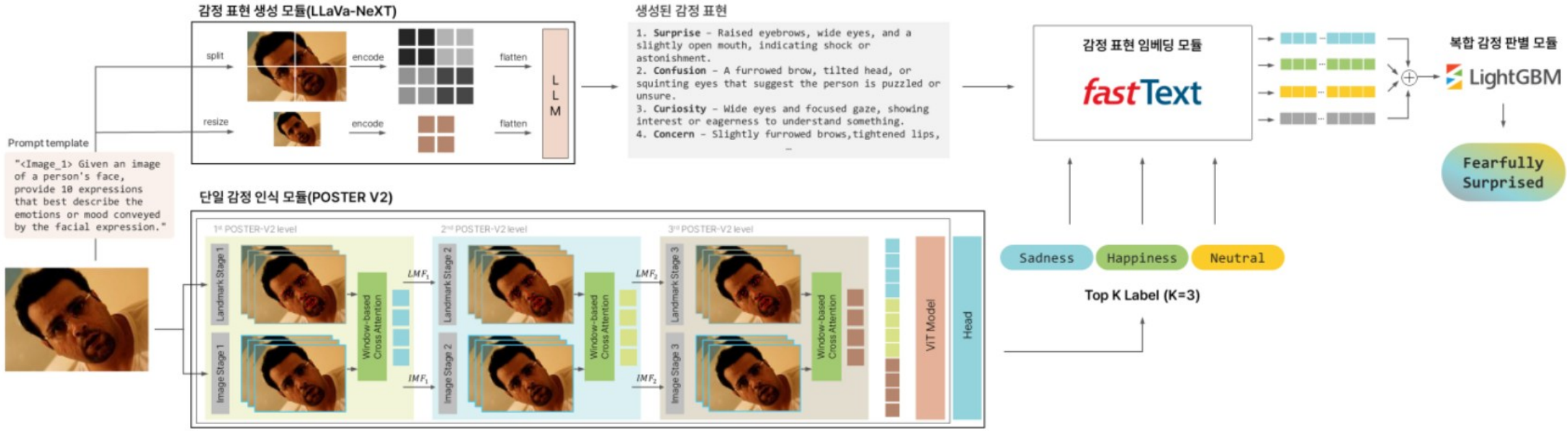
- fastText 임베딩 모델을 사용하여 선택된 단일 감정 라벨들과 VSM 모델이 생성한 감정 표현을 벡터로 변환하여, 복합 감정 판별을 위한 입력으로 활용

복합 감정 판별 모듈

- LightGBM을 사용하여 텍스트 임베딩 모듈에서 생성된 벡터를 입력으로 받아 복합적인 얼굴 감정 라벨을 예측

VSM 모델을 통한 복합 감정 인식 파이프라인

데이터셋: RAF-DB(11가지 복합 감정 라벨-3954개의 이미지)



결론

표 1. RAF-DB 복합 감정 인식 정확도 (척도: UAR)

RAF-DB (CE)	UAR
POSTER-V2 + LLaVa-NeXT	**0.496**
Fine-tuned FaceBehaviorNet [10]	0.483
VGG + mSVM [9]	0.316
baseDCNN + mSVM [9]	0.402
DLP-CNN + mSVM [9]	0.446
ResNet 18 + separate loss [11]	0.432
ReCNN [12]	0.461
ResNet-18 (ARM) [13]	0.471
PSR [14]	0.465
DACL [15]	0.466

- 비전-언어 통합을 통한 복합 얼굴 감정 인식 방법을 제안

- VSM 모델 라벨의 활용이 성능에 기여했으나, 개선이 한계에 도달하는 지점도 확인

Thank you