사진 데이터로 본 미세먼지 단계 추정 시스템 : 딥러닝 기술의 적용

박현지[1], 정지영[2], 김유정[3], 박현수[4], 최현지[5]
[1,2]덕성여자대학교 컴퓨터공학과 학부생
[3,4,5]덕성여자대학교 컴퓨터공학전공 학부생
phjgina0314@gmail.com, gzero21@naver.com, kimyj4017@gmail.com,
dontlikesm@gmail.com, ch5i_hj15@naver.com

Estimation of Fine Dust Concentration Using Photo Data : Application of Deep Learning

Hyeon-Ji Park[1], Ji-Young Jeong[2], Yu-Jung Kim[3],
Hyun-Soo Park[4], Hyun-Ji Choi[5]
[1,2,3,4,5]Dept. of Computer Engineering, Duksung Women's University

요 약

미세먼지 단계를 예측하는 딥러닝 기반 시스템을 개발하고 그 성능을 평가하는 연구를 진행했다. 연구에서 320개의 풍경 사진 데이터를 수집하고, 해당 시점의 미세먼지 농도를 측정하여 "좋음" 또는 "나쁨"으로 분류했다. 데이터 전처리 단계에서는 특히 하늘 이미지의 특성을 고려하여 다양한 전처리 기법을 적용하였다. 다섯 가지 이미지 데이터 모델을 사용하여 이미지를 분류하고 미세먼지 단계를 예측하는 모델을 개발하였으며, 또 이 모델들을 다양한 기법으로 앙상블 해보며 성능을 비교했다. 그 결과, Random Forest를 이용한 앙상블 모델이 제일 뛰어난 예측 성능을 보였다. 이러한 연구 결과는 미세먼지 모니터링 및 예측에 유용한 시스템 개발의 가능성을 제시한다.

1. 서론

미세먼지는 현대 도시에서 심각한 환경 문제 중 하나로, 인간의 건강 및 환경에 부정적인 영향을 미치는 주요 대기 오염물질이다. 따라서 미세먼지의 농도와 분포를 정확하게 모니터링하고 예측하는 데는 실시간 정보가 필요하며, 이를 위해 현대적인 기술 및 방법이 요구된다.

본 연구는 미세먼지 농도를 평가하기 위해 멀리 있는 건물을 관찰하는 사람들의 모습에서 착안하여, 딥러닝 기술을 활용해 풍경 사진을 분석하고, 미세먼지 단계를 추정할 수 있는 시스템을 개발하는 것을 목표로 한다.

딥러닝의 뛰어난 시각 데이터 처리 능력은 미세먼지 추정을 위한 새로운 가능성을 열어주며, 이를 통해 미세먼지를 효율적으로 모니터링하고 환경에 관한 정보를 제공받을 수 있을 것이다.

2. 연구 방법

2-1. 데이터 수집

6~8월간 직접 사진을 찍어 320개의 데이터를 수집했다. 하늘 이미지가 반드시 영상에 포함되도록 했으며, 수집할 당시 미세먼지 농도를 측정하여 데이터 분류에 사용했다.

데이터 분류는 환경부가 제시한 기준으로 미세먼지 농도가 좋음과 보통(0~35μg/㎥)에 해당할 경우, good, 나쁨과 매우 나쁨(36μg/㎥ 이상)에 해당한 경우, bad로 분류하였다. 위 기준으로 분류하여 good에 해당하는 데이터는 190개, bad에 해당하는 데이터는 130개로 연구를 진행하였다.

2-2. 데이터 전처리

과적합을 방지하기 위해, 수집한 사진 데이터들을 학습시키기 전, 데이터 증강을 위한 다양한 변형을 통한 전처리 과정을 거치도록 했다.

본 연구에서 다루는 데이터인 풍경(하늘) 사진은 특징적인 패턴과 구조가 부족하므로, 과도한 변형은 오히려 모델의 학습을 방해할 수 있다. 따라서 여러 번의 테스트 후 정확도를 고려하여 RGB 값 재조정, 이미지 회전, 이동, 확대 및 축소, 뒤집기와 빈 공간 보정을 이용하여 데이터 전처리를 진행하였다.

2-3. 이미지 데이터 모델 소개

1) RGB 모델: RGB 이미지를 입력으로 사용하며 일반적인 컬러 이미지 처리 기술을 활용한다.

2) LAB 모델: 이미지를 LAB 색 공간으로 변환하여 색상 정보를 활용하는 모델이다.

3) Dark Channel 모델: 이미지의 가장 어두운 채널 정보를 추출한다.

4) Deep Dust 모델: 'gaussian_blur_residual'을 적용하여 이미지의 높은 주파수 성분을 줄여 잔상 이미지를 계산, 최종적으로 원본에서 가우시안 블러를 적용한 이미지를 뺀 이미지를 반환하여 노이즈(먼지)를 제거한다.

5) Haze Removal 모델: 데이터를 YUV로 변환 후 어두운 채널과 대기광을 계산한다. 이를 이용하여 어두운 채널에서의 대기광과 해당 픽셀의 밝기 간의 관계(전달 함수)를 계산한다. 이 함숫값이 안개의 정도를 뜻하며, 최종적으로 이미지의 안개가 제거된 이미지를 반환한다.

2-4. 학습 및 평가 방법

각 이미지 데이터 모델은 학습 데이터 셋을 사용하여 학습되었으며, 손실 함수와 최적화 알고리즘이 설정되었다. 또한, 학습 과정에서 Early Stopping을 사용하여 과적합을 방지하려 했다.

이미지 데이터 모델의 학습 및 검증 과정에서 손실과 정확도를 모니터링했고, 각 모델의 성능을 평가하였다. 학습 및 검증 결과는 그래프와 표로 시각화하여 손실 및 정확도의 변화를 확인했다.

학습된 다섯 가지 모델의 예측 확률을 결합하기 위해 Soft Voting 앙상블과 Random Forest 앙상블을 개별적으로 구현하여, 최종 예측을 생성하고 정확도, 정밀도, 재현율, F1 점수를 측정하여 모델의 종합 성능을 평가, 비교했다.

3. 결과

	RGB	LAB	Dark Channel	Deep Dust	Haze Removal	**Soft Voting**	**Random Forest**
정확도	0.5149	0.4653	0.4554	0.4356	0.4752	**0.4993**	**0.6238**
정밀도	0.5000	0.4603	0.4375	0.4375	0.4667	**0.4528**	**0.6222**
재현율	0.6122	0.5918	0.4286	0.5714	0.5714	**0.4897**	**0.5714**
F1 점수	0.5505	0.5179	0.4330	0.4956	0.5138	**0.4706**	**0.5957**

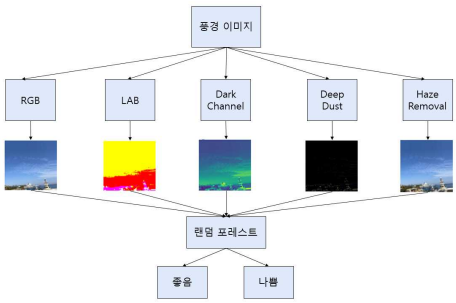
<표 1> 학습 결과

이미지 데이터 모델 중에서는 RGB 모델이 가장 높은 정확도와 재현율을 가지고 있으며, F1 점수 역시 상당히 높다. LAB 모델과 Haze Removal 모델도 비교적 좋은 성능을 가지고 있음을 알 수 있다.

앙상블 모델 중에서는 Random Forest 모델이 개별 이미지 데이터 모델들뿐만 아니라 Soft Voting 모델에 비해서도 정확도와 정밀도가 높다. 또한, 정밀도와 F1 점수가 향상된 것을 볼 수 있는데, 이를 통해 Random Forest 모델은 다양한 개별 이미지 데이터 모델의 예측을 조합하여 전반적으로 더 나은 예측 성능에 도달한 것을 알 수 있다.

4. 결론 및 논의

본 연구 결과를 종합해, 다양한 이미지 데이터 모델을 조합한 앙상블 모델 중 뛰어난 예측 성능을 보여준 Random Forest 모델을 최종 모델로 제안하고자 한다. Soft Voting 모델의 경우, 5가지 이미지 데이터 모델의 예측 확률값을 단순 평균 내어 최종 예측값을 도출했지만, 5가지 각각의 모델이 높은 정확도를 보여주지 않았기에 Random Forest 방식을 사용하는 게 더 합리적이라 판단하였다.



(그림 1) 최종 제안 모델

추후 연구에서는 본 연구에서 사용한 앙상블 모델 외에도 다양한 분류 모델(bagging, boosting 등)을 사용하는 등 다양한 분류 방법을 위한 전처리 및 분류 모델의 성능을 비교하여 보다 예측 성능이 뛰어난 모델을 찾아내는 것이 계획되어 있다.

이와 더불어 데이터 셋 확장 및 다양한 환경 조건에서의 테스트 등 다양한 연구를 고려한다면, 사진을 이용하여 미세먼지 단계를 예측할 수 있다는 점에서 실제 환경에서 미세먼지 모니터링 및 예측에 유용하게 활용될 수 있을 뿐만 아니라 더 나은 예측 성능을 기대할 수 있을 것이다.

참고문헌

[1] 김송이. “딥러닝에 기반한 미세먼지 예측 통합 모델.” 국내석사학위논문 계명대학교, 2019. 대구

[2] 공도경 외. “하늘 이미지를 활용한 미세먼지 수치 추정”, 날씨 빅데이터 콘테스트, 기상청, 2018

※본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.