

도메인 일반화를 위한 프롬프트 앙상블 기반 Zero-shot 화재 탐지 기법

정성균, 최흥준, 이민호, 이원일, 이수빈, 유연승, 김태형*

Seonggyun.Jeong@lotte.net, heungjunchoi@lotte.net, minho_lee@lotte.net, wonil.lee@lotte.net, leesubin@lotte.net, yys4000@lotte.net, thkim07@lotte.net*

롯데이노베이트(주)

Prompt Ensemble-based Zero-shot Fire Detection for Domain Generalization

Jeong Seonggyun, Choi Heungjun, Minho Lee, Wonil Lee, Lee Subin, Yu Yeon Seung, Kim Taehyung*

Seonggyun.Jeong@lotte.net, heungjunchoi@lotte.net, minho_lee@lotte.net, wonil.lee@lotte.net, leesubin@lotte.net, yys4000@lotte.net, thkim07@lotte.net*

LOTTE INNOVATE Co.,Ltd.

요약

최근 증가하는 화재 사고를 방지하기 위해 다양한 딥러닝 모델이 제안됐으나 기존 지도 학습 방식의 모델의 경우 일반화 부족으로 환경이 바뀌었을 때 성능이 감소할 수 있는 문제가 존재한다. 이를 해결하기 위해 사전 학습된 텍스트-이미지 모델인 CLIP과 프롬프트 앙상블(prompt ensemble) 방식을 결합하여 zero-shot 화재 탐지 구조를 본 논문에서 제안한다. 해당 방식의 경우 기존 방식에 비해 비교적 작은 연산량과 더불어 효과적인 탐지 성능을 보이고, 실시간 화재 감지 시스템에 적극적으로 활용될 수 있다.

I. 서론

최근 국내 발생 화재 건수는 전년 대비 약 10% 증가한 3,444 건 [1]이며 화재 사건은 사회적·경제적 손실을 초래하기에 해결이 시급한 문제이다.

해당 문제를 해결하기 위해서 규칙 기반의 실시간 화재 정보 시스템이 개발되었지만, 이 시스템의 경우 규칙의 일반화가 다소 어려워 특정 환경에서는 정확도가 감소하는 모습을 보인다. 따라서 다양한 환경에서도 신뢰성 있는 결과를 출력하는 시스템이 필요하며 이에 따라 딥러닝 기반의 영상 화재 정보 시스템이 최근 활발히 개발이 진행되고 있다. 주로 실제 화재 이미지를 수집 후 정답을 만들어 지도 학습 방식으로 이를 구현한다. 그러나 이러한 방식은 수집된 데이터가 특정 지역 혹은 국가에 국한되기에 해당 특성과 다른 도메인의 데이터가 입력될 경우에 여전히 오 예측이 발생할 가능성이 존재한다.

따라서, 도메인의 일반화를 위해서 모든 데이터를 수집하는 것이 아니라, 기존의 사전 학습된 모델을 이용해 재학습 없이 일관성 있는 예측을 하는 zero-shot 화재 탐지 모델 [2], [3]이 필요로 해 최근 논문에서 제안됐다. 해당 구조는 CLIP [4]이라는 구조를 이용해서 zero-shot 화재 탐지를 구현했는데, CLIP은 텍스트-이미지 기반의 모델로서 대용량 사전 데이터로부터 이미지와 텍스트 간의 유사도를 출력하도록 사전 학습됐기에, 다양한 도메인의 화재 이미지를 탐지할 수 있다. 구체적으로 화재 이미지가 들어왔을 때 화재 관련 텍스트와 비화재 텍스트의 유사도를 비교해서 판별하는 것이며, 해당 방식을 통해 다양한 데이터 셋에서 효과적인 탐지 성능을 보였다.

하지만 여러 개의 화재 및 비화재 프롬프트 각각에 대한 유사도를 계산하므로 계산량이 증가하며 통합된 프롬프트를 사용하면 성능이 더 향상될 여지가 존재한다.

따라서 본 연구에서는 통합된 프롬프트를 사용해 zero-shot 기반의 화재 탐지 구조를 제안한다. 구체적으로 프롬프트 앙상블 방식을 통해 프롬프트를 통합했으며, 여러 프롬프트의 특성 벡터를 중합해 평균 화재 벡터와 비화재 벡터를 구해 이에 대한 유사도로 화재를 탐지한다. 본 방식은 여러 번의 유사도 계산을 하지 않으므로 비교적 적은 연산량을 가지면서도 향상된 탐지 성능을 보인다.

실험에 사용된 모델은 별도의 재학습 없이, CLIP ViT-B/16 [5]의 이미지 인코더와 텍스트 인코더의 사전 학습된 가중치를 그대로 활용했으며, 텍스트 프롬프트는 표 1과 같이 기존 연구 [2], [3]에서 제안된 문장을 사용했다.

표 1 화재, 비화재 프롬프트

화재 프롬프트
There is no person. There is a moving person. There is a person doesn't move.
비화재 프롬프트
There is smoke rising. There is an arsonist. There is something that shines brightly. There is a campfire. There is a flame and fire soaring.

II. 본론

본 논문에서 제안하는 화재 탐지 모델의 전체 구조는 **그림 1** 과 같으며 우선, 추론 대상 이미지를 이미지 인코더에 입력하여 이미지 임베딩 벡터 v_1^i 를 얻는다.

이어서, 사전에 정의된 화재 관련 텍스트 프롬프트와 비화재 관련 텍스트 프롬프트 각각 텍스트 인코더에 입력하여, 화재 프롬프트에 대한 임베딩 벡터 집합 $[v_1^n, v_2^n, \dots, v_n^n]$ 과 비화재 프롬프트에 대한 임베딩 벡터 집합 $[v_1^f, v_2^f, \dots, v_n^f]$ 를 생성한다. 기존 방식 [2], [3]에선 이미지 임베딩 벡터와 화재, 비화재 프롬프트 임베딩 집합의 원소에 대해 모두 유사도를 계산해서 가장 높은 유사도를 보이는 3 개의 텍스트 프롬프트 모두 화재 프롬프트에 속해야 화재로 판별한다.

하지만 본 연구에서 각 텍스트 임베딩 벡터 집합의 전체적인 의미를 통합적으로 반영하기 위해, **그림 1** 과 같이 유사도 모두 계산하는 것이 아니라 각각의 벡터 집합에 대해 평균 연산을 수행하여 화재 평균 임베딩 벡터 v^n , 비화재 평균 임베딩 벡터 v^f 를 계산한다. 그다음, 이미지 임베딩 벡터 v_1^i 와 평균 텍스트 임베딩 벡터들 간의 유사도를 계산하기 위해 매트릭스 곱을 수행한다. 즉, $v_1^i \cdot v^n$ 는 화재 점수, $v_1^i \cdot v^f$ 는 비화재 점수가 된다. 각 점수는 0 과 1 사이의 값으로 표현되며 두 점수를 비교해, 화재 점수가 더 높은 경우에 화재로 분류한다.

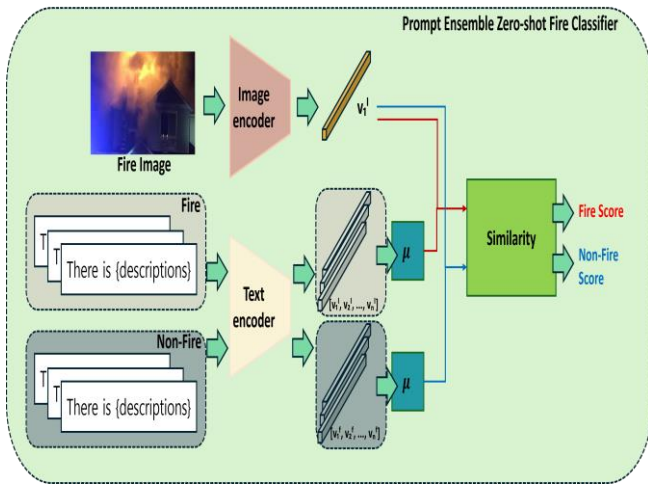


그림 1 제안하는 화재 탐지 파이프라인

III. 실험

제안하는 구조의 화재 탐지 성능을 검증하기 위해, FireNet [6] 데이터셋을 활용하였다. FireNet 은 다양한 해상도의 화재 이미지로 구성되어 있으며, 훈련은 412 장, 검증용 90 장의 이미지를 포함하고 있다. 모든 이미지는 이미지 인코더에 입력하기 위해 224×224 크기로 리사이즈 한 후 사용하였다.

표 2 은 FireNet 데이터셋에서의 실험 결과를 정리한 것이며 Zero-shot 분류 성능 비교를 위해, 기존 텍스트 프롬프트 기반 방식 추론이 가능한 YOLOE [7]을 비교군으로 추가했으며, 구체적으로 YOLOE-11-L 가중치를 사용하였다.

또한, CLIP 모델의 경우 패치 크기에 따라 성능 차이가 발생하므로, 본 모델 (CLIP ViT-Base/16) 외에도 CLIP ViT-Base/32 를 추가로 비교하였다. 사용된 그래픽 카드의 경우 NVIDIA H100 80GB 이며 추론 배치는 1 로 고정해서 FPS 를 측정하였다. Accuracy, F1 score 는 1 에 가까울수록 모델의 예측 결과가 정답에 유사함을

의미하며, FPS 는 초당 추론한 이미지 개수로 값이 클수록 추론속도가 빠름을 의미한다. 제안하는 프롬프트 앙상블 기반 구조가 정확도 0.9976, F1 score 0.998, FPS 96.24 로 가장 우수한 성능과 빠른 추론속도를 기록하였다.

표 2 FireNet zero-shot 성능 비교

모델	Accuracy	F1 Score	FPS (Img/s)
YOLOE-11-L	0.313	0.476	22.73
CLIP (ViT-B/32)	0.895	0.944	85.47
CLIP (ViT-B/16)	0.946	0.972	64.1
+ Prompt Ensemble	0.997	0.998	95.24

IV. 결론

본 연구를 통해서 기존의 텍스트-이미지 모델을 활용한 화재 탐지 모델의 탐지 성능을 향상시킬 뿐만 아니라 추론 속도 또한 효율적으로 높일 수 있었다.

제안한 방법은 텍스트 프롬프트를 종합하는 방식이기에 다양한 구조에 손쉽게 결합될 수 있으며, 기존의 화재 방지 CCTV 시스템에 적극적으로 도입해 화재 사건 방지에 기여할 수 있다.

현재 모델을 단순 화재와 방화 사건을 구별하지는 못하므로 추후엔 텍스트 프롬프트와 구조의 고도화를 통해서 화재만 판단하는 것이 아닌 원인 분석까지 가능한 파이프라인 개발로 확장할 예정이다.

참 고 문 헌

- [1] 소방청, 화재통계, 2025, (<https://www.nfds.go.kr/stat/general.do>).
- [2] Radford Alec, "Learning transferable visual models from natural language supervision." International conference on machine learning.
- [3] Jeon, Hobeom, "Zero-shot Fire And Arson Detection Using Textual Descriptions." *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2022.
- [4] Jeon Hobeom, "PASS-CCTV: Proactive Anomaly surveillance system for CCTV footage analysis in adverse environmental conditions.", *Expert Systems with Applications* 254, 2024.
- [5] Dosovitskiy Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929*, 2020.
- [6] Jadon Arpit, "FireNet: a specialized lightweight fire & smoke detection model for real-time IoT applications.", *arXiv preprint arXiv:1905.11922*, 2019.
- [7] Wang Ao, "Yoloe: Real-time seeing anything.", *arXiv preprint arXiv:2503.07465*, 2025.