Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model

Yong, G., Jeon, K., Gil, D., & Lee, G. (2022). Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Computer-Aided Civil and Infrastructure Engineering.

Introduction

- 과거부터 AI를 활용하여 결함을 탐지하는 연구가 진행됨
- 하지만 결함은 자주 발생하는 것이 아니기에 결함에 대한 instance는 항상 적음
- 작은 결함 데이터셋으로도 좋은 성능의 결함 탐지 모델을 만들기 위해
  - GAN 같은 augmentation algorithm을 활용해 결함 데이터셋을 추가로 생성
  - Transfer Learning을 활용
  - 대규모 language model을 활용한 Few-shot / Zero-shot learning 활용
- vision-language pretrained (VLP) model은 customization 없이 다양한 작업을 수행하도록 개발
  - 대규모의 이미지와 언어 데이터셋을 이용해 ImageBERT, UNITER 등 다양한 모델 개발
  - 그 중 CLIP model은 간단한 모델 구조와 강력한 zero-shot 성능을 보임
  - 하지만 독일 교통 신호 인식 벤치마크(GTSRB) 작업, 림프절 종양 탐지(Patch Camelyon) 등 전문적인 분야에서는 여전히 낮은 성능을 보임

Introduction

- VLP model이 개발되면서 이미지에 대한 prompt 설정에 대한 연구 또한 진행
  - CLIP, Dall-E, Midjourney 와 같은 모델들은 prompt에 따라 성능 변화가 큼[1]
  - CLIP 논문에서는 “a {category} photo of {label}” 형태의 prompt 사용
  - 하지만, 추후 연구[2]에서 prompt를 random or pre-trained word embedding으로 learnable vector를 초기화 하고, class와 함께 text encoder를 학습시키는 것이 더 좋은 성능을 보이는 것을 보임
- 본 논문은 CLIP의 zero-shot 기법을 활용한 결함 탐지에서 좋은 성능을 보이는 prompt의 특징을 파악하고자 함

1. Nine, A. (2022). People have begun to sell their prompts for AI-generated artwork. https://www.extremetech.com/internet/ 339304-people-have-begun-to-sell-their-prompts-for-ai-generated-artwork
2. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337-2348.

Defect Detection with Insufficient Dataset

결함 데이터셋 부족 문제와 연구 사례

- Transfer Learning
  - Y. Gao와 Mosalam (2018): 1600개 이미지를 사용해 VGG-16 훈련[1]
  - Liang (2019): AlexNet, Google Net, VGG-16 등의 모델에 1154개 이미지로 훈련 [2]
  - S. Jiang과 Zhang (2020): SSDLite-MobileNetV2를 사용해 1030개 이미지로 훈련 [3]

- Data Augmentation
  - J. Zhu et al. (2020): 밝기, 채도 조절, Flip을 사용해 243개 원본 결함 이미지를 1458개로 증강 [4]
  - Y. Li et al. (2021): GAN을 사용해 2500, 5000, 7500, 10000개로 증강된 포장 도로 불량 이미지로 훈련 [5]
  - Maeda et al. (2021): GAN을 사용해 1200개 원본 이미지를 1800, 2400개로 증강 => SSD MobileNet으로 학습 [6]

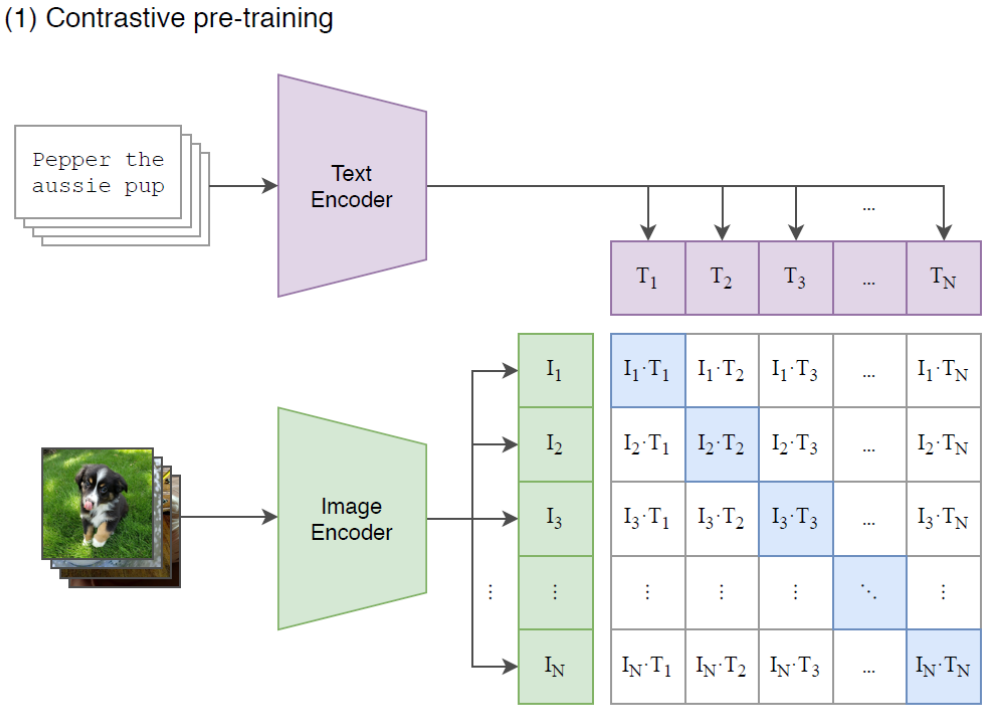
Defect Detection with Insufficient Dataset

불균형 데이터셋 해결 방법과 연구 사례

- Oversampling
  - Meijer et al. (2019): 클래스 가중 손실 함수를 적용해 17,663개 이미지를 5배 확장 [7]
- Meta-learning
  - Guo et al. (2020): 메타러닝 기반 CNN으로 21,259개 이미지를 사용해 건물 외관 결함 분류[8]
- Semi-supervised Learning
  - Guo et al. (2021): CNN 기반 Semi-supervised Learning으로 총 5621개 이미지로 건물 외관 결함 분류 [9]
  - Y. Gao et al. (2021): GAN을 통한 오버샘플링과 Semi-supervised Learning으로 10,500개 이미지로 훈련 [10]
- Zero-shot & Few-shot Learning
  - Cui et al. (2022): 1-shot, 2-shot, 5-shot, 10-shot 방법을 사용해 건물 외관 결함 분류 [11]

CLIP[1]

- 자연어 지도 학습을 통한 시각적 표현을 학습하는 vision-language pretrained (VLP) model
- 텍스트 인코더와 이미지 인코더가 짝이 맞는 텍스트, 이미지를 찾을 수 있도록 학습
- 다양한 시각적-언어적 작업을 위한 Zero-shot 및 Few-shot 능력을 지님



1. Radford, A., ... (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

CLIP[1]

학습 방법

1. 전처리: 이미지 및 텍스트 데이터 수집 및 처리

2. 이미지-텍스트 쌍 생성: 이미지와 설명 쌍을 만듦

3. 학습 : Contrastive Learning을 사용하여 이미지와 텍스트를 embedding space로 mapping
   - 이미지 인코더: 이미지를 고차원 특성 벡터로 변환 (예: ResNet, ViT 등)
   - 텍스트 인코더: 텍스트를 고차원 특성 벡터로 변환 (예: BERT, GPT 등)

4. 최적화: 코사인 유사도를 사용하여 이미지와 텍스트 임베딩 간의 거리를 최소화

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

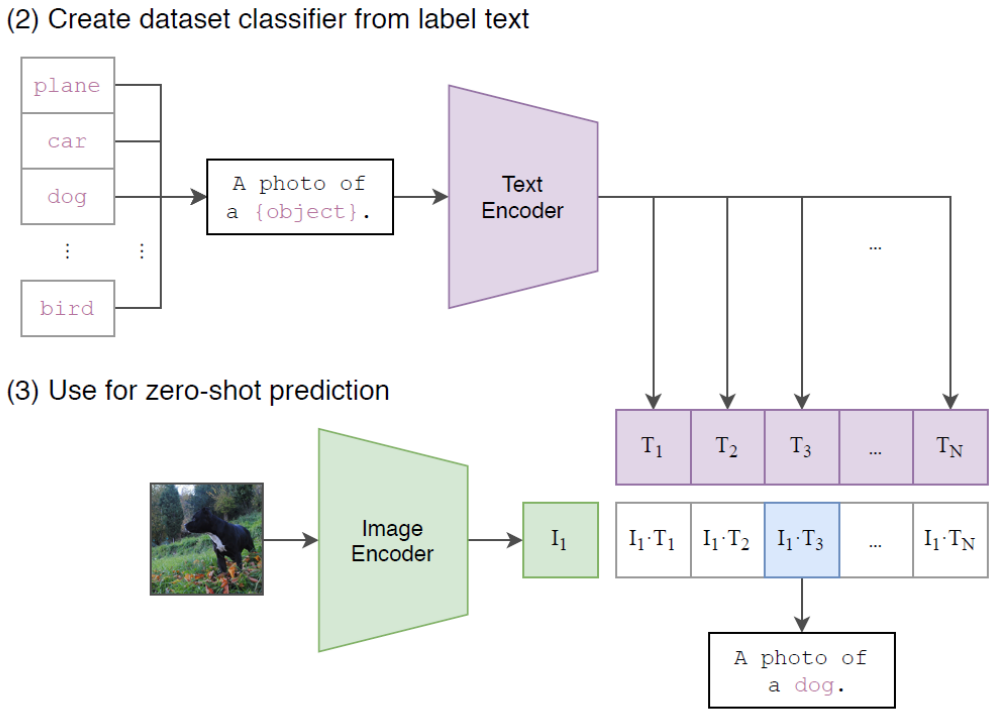
Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

1. Radford, A., ... (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

CLIP[1]

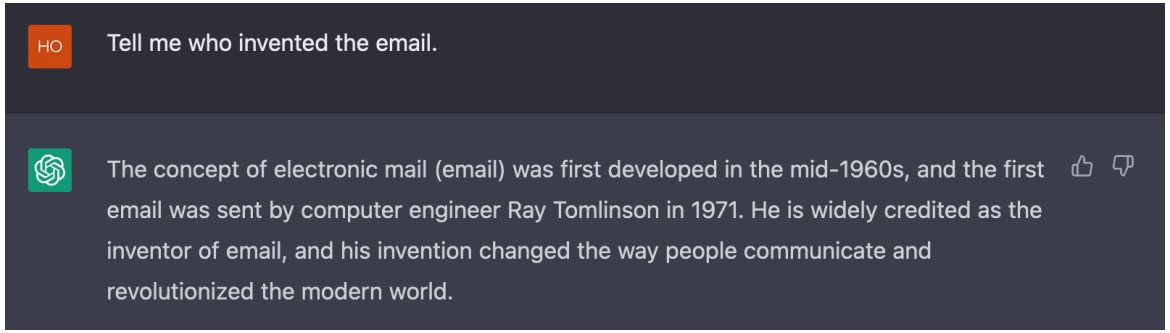
CLIP 모델을 활용한 제로샷 방법 및 성능

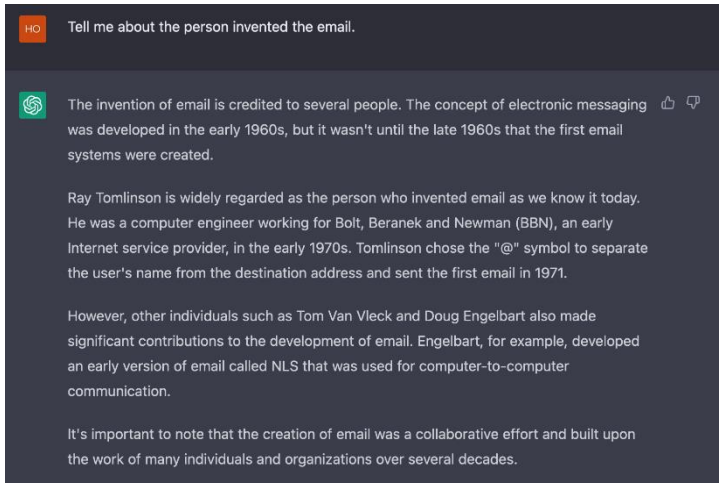
- 제로샷 학습: 사전에 학습된 모델을 새로운 작업에 즉시 적용
- 성능: 다양한 시각적-언어적 작업에서 기존 접근 방식보다 우수한 성능을 보임
  - 이미지 분류, 객체 인식, 자연어 처리 작업 등에서 좋은 결과
- 제로샷 성능 비교:
  - 고전적인 전이 학습 방법보다 더 나은 결과를 보임
  - 더 적은 양의 라벨링 데이터로도 높은 성능을 달성 가능
- CLIP 모델은 미래의 다양한 시각적-언어적 작업에 유용한 기반이 될 것으로 기대



Prompt

- 최근 Chat-GPT, BARD 같은 대화형 인공지능(Conversation AI) 등장
- 아직 완벽한 단계가 아니기에 질문의 intent 파악을 못하거나 잘못된 정보를 return 해줌
- 거대 언어 모델로부터 높은 품질의 응답을 얻어낼 수 있는 프롬프트 입력 값들의 조합을 찾을 필요가 있음(GIGO)
  - 프롬프트(Prompt) : 거대 언어 모델(Large Language Model; LLM)로부터 응답을 생성하기 위한 입력 값
  - GIGO(garbage in garbage out) : 무가치한 데이터를 넣으면 이상한 결과가 나온다





<프롬프트 문구들의 미세한 차이에 따른 결과물 변화 예시>

Prompt Engineering

- Prompt Engineering
  - 입력 쿼리 재구성으로 CLIP 같은 사전학습 모델 성능 향상
  - 목표 정보를 높은 정확도로 식별할 수 있게 쿼리 수정

- Prompt Tunning의 한계
  - 결과를 미리 알아야 최적화 가능
  - 초기 프롬프트 구성 방법이 미흡

- 본 연구의 차별점
  - "프롬프트 튜닝"이 아닌 "초기 프롬프트 구성"에 초점을 맞춤

PROPOSED PROMPT ENGINEERING METHOD

본 논문에서 제안하는 PROMPT 구성에 대한 가설은 다음과 같음

- (H1): A DK based definition of a defect performs better as a prompt than a GK based definition
  - DK : 도메인 지식을 활용한 prompt / GK : 일반적으로 사용하는 용어들로 표현한 prompt / BL : CLIP에서 제안하는 기본 prompt
  - 도메인 특화 사전(DK 프롬프트)과 일반 사전(GK 프롬프트)을 사용한 프롬프트 간 차이 분석
  - 또 다른 실험으로 임베딩한 DK, GK prompt를 앙상블 하여(Ex, DK1~DK3 =>DK_ensemble) 정보를 앙상블한 효과를 검증

Mildew (mold; mould)
[BL] A defect photo of mildew (R1)
[DK1] Mildew is a fungus growth that is enhanced by dampness (R2)
[DK2] Mildew is a fungus that grows and feeds on paint, cotton, and linen fabric, and so forth, which are exposed moisture; causes discoloration and decompositions of the surface (R3)
[DK3] Mildew is a fungus that stains materials but does not rot wood (R8)
[GK1] Mildew is a woolly, furry, or staining growth now recognized as consisting of fungus, such as that which forms on food, textile, and so forth (R5)
[GK2] Mildew is a fungus producing mildew (R6)
[GK3] Mildew is a white or gray substance that grows on walls or other surfaces in wet, slightly warm conditions (R7)

<Defection에 대한 prompt 예시>

TABLE 2 Prompt source identifiers and references

ID	Reference
R1	Learning transferable visual models from natural language supervision (Radford et al., 2021)
R2	Dictionary of Construction Terms (Tolson, 2012)
R3	Dictionary of Architecture and Construction (Harris, 2006)
R4	Dictionary of Building and Civil Engineering (Montague, 2017)
R5	Oxford English Dictionary (Simpson & Weiner, 1989)
R6	Dictionary of Merriam-Webster (Merriam-Webster, 2019)
R7	Longman Dictionary of Contemporary English (Pearson Education, 2014)
R8	Dictionary of Building (Scott & Maclean, 2000)
R9	A professional website (InspectApedia, n.d.)
R10	Crest Wood Painting (Crestwoodpainting, n.d.)
R11	A Dictionary of Construction, Surveying and Civil Engineering (Gorse et al., 2012)
R12	National Dictionary of Building & Plumbing Terms (Australia; Standards Australia, n.d.)

< prompt 출처>

PROPOSED PROMPT ENGINEERING METHOD

본 논문에서 제안하는 PROMPT 구성에 대한 가설은 다음과 같음

- (H2): A list of core terms in a definition performs better as a prompt than a complete sentence definition
  - 불용어(Stop word) : 분석에 큰 의미가 없는 단어
  - VLP 모델에서 stop word를 제거하는 것이 좋은지, 허용한다면 얼마만큼 넣어야 하는지 테스트 진행

- (H3): A defect image is better than the defect definition as a prompt.
  - 시각(이미지) vs 텍스트(프롬프트) 정보의 설명력 비교 측정 진행
  - 동일한 클래스의 여러 개의 이미지를 임베딩한 벡터의 평균을 계산하여 해당 클래스의 평균적인 시각 정보 영향을 측정할 수 있음
  - 해당 실험을 통해 결함 분류나 감지를 위한 프롬프트 최적화가 가능 & VLP 모델에 효과적으로 전달할 수 있는 형식 파악 가능
  - 또한, 여러 개의 이미지를 임베딩한 벡터의 평균을 이용하기에 “a few-shot transfer”와 유사하다고 말할 수 있음

PROPOSED PROMPT ENGINEERING METHOD

본 논문에서 제안하는 PROMPT 구성에 대한 가설은 다음과 같음

- (H4): A multimodal prompt with the combination of defect images and definitions performs better than a single-modal prompt
  - VLP 모델이 정보를 처리할 때 결함 이미지와 프롬프트 사이의 간격이 존재하는지 확인하기 위해 PCA(주성분 분석)를 수행
  - VLP 모델이 input 정보를 분석할 때, data type에 따른 차이 분석
  - [1]은 텍스트와 이미지 정보를 결합한 멀티 모달 방식이 싱글 모달에서 실패한 새로운 언어적 또는 시각적 개념을 확인할 수 있었음
  - [1]과 같이 앙상블 공간에서 통합하여 VLP 성능을 최대화할 수 있는 prompt 분석

1. Shibata, T., Kato, N., & Kurohashi, S. (2007, September). Automatic object model acquisition and object recognition by integrating linguistic and visual information. In Proceedings of the 15th ACM international conference on Multimedia (pp. 383-392).

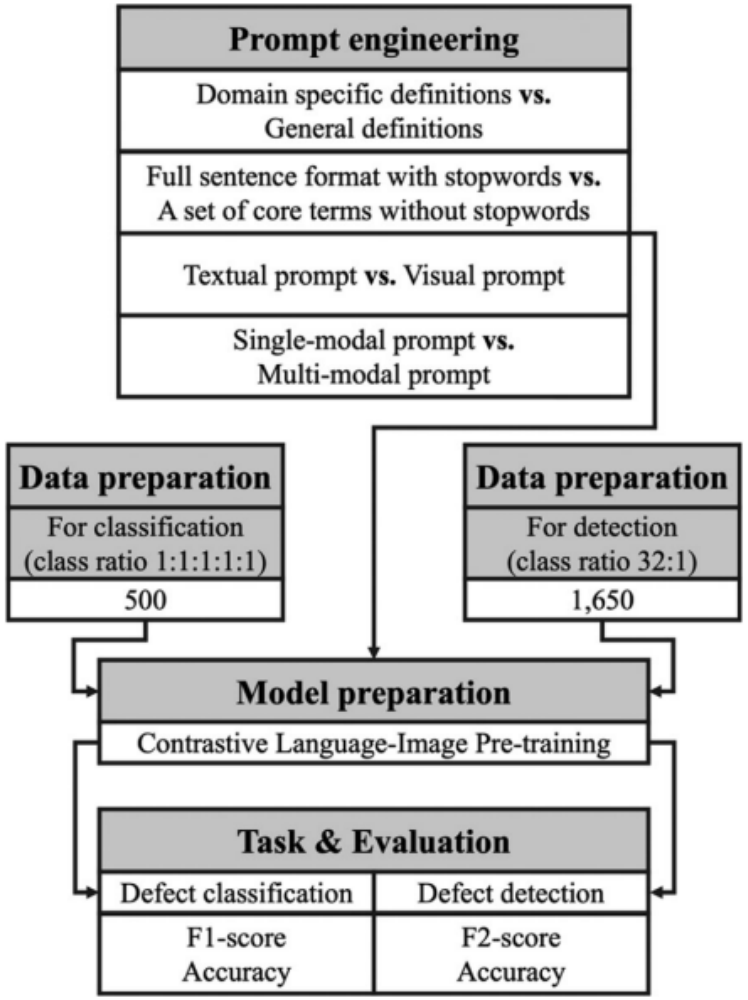
EXPERIMENT DESIGN

- Image Detection vs Image Classification
  - Classification : 이미지를 class별로 분류하는 것 (어떤 결함인지)
  - Detection : image batch에서 defection 이미지를 찾아내는 것 (결함인지 아닌지)

- Performance indices
  - F_β score : recall 혹은 precision에 대해 β 만큼 가중치를 부여하여 구하는 Precision과 Recall의 조화 평균 값
  - 본 논문에서는 β = 2 로 설정(Recall에 더 비중)

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

- Dataset
  - Dataset은 다음 그림과 같이 5 class
  - Zero-shot 이기에 train dataset X
  - Defection : 1600개의 정상 데이터 & 50개의 defection image(클래스당 10개)
  - Classification : 4800개의 정상 데이터 & 300개의 defection image(클래스당 60개)



GURE 2 Methodology of the present study

RESULTS AND ANALYSIS

Domain knowledge prompts versus general knowledge prompts

- Classification에서는
  - 도메인 지식 기반(DK)이 일반적인 지식 기반(GK)보다 좋은 성능을 보임
  - 프롬프트를 조합하면 성능이 일반적으로 개선됨
  - DK를 사용하는 분류에서 "peeling“ class를 가장 분류 못함

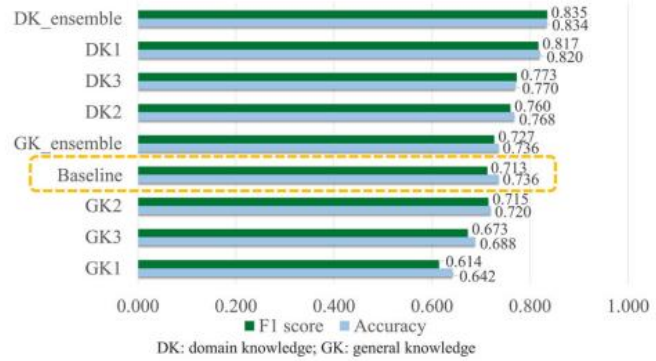


FIGURE 4 Zero-shot defect classification performances

- Detection에서는
  - DK & GK 모두 baseline보다 좋은 성능을 보임
  - GK_ensemble & DK_ensemble 은 거의 유사한 성능을 보임
  - T-test 결과, DK와 GK 프롬프트 간에 통계적으로 유의미한 성능 차이가 없음

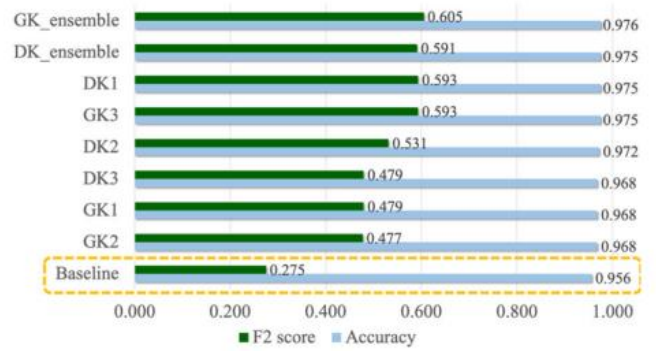


FIGURE 5 Zero-shot defect detection performances

RESULTS AND ANALYSIS

Complete sentence definitions against core terms

- 대부분 경우에서 Stop words를 제거했을 때 성능 저하

- Stop words 가 defection에 대한 핵심 정보를 전달하지 못할 수도 있지만 이미지 분류 및 탐지에 중요한 역할 수행
  - Stop words는 VLP 모델에 핵심 용어 간의 맥락적 관계를 제공하기 때문에 성능 향상에 기여한다 추측

- Spearman's rank correlation coefficient 결과, 프롬프트의 단어 수는 zero-shot classification에 큰 영향 X
  - prompt 구성에 따라 성능 변화

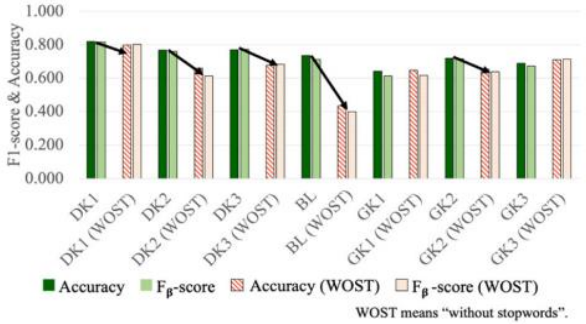


FIGURE 6 Defect classification performance change after the removal of stopwords in a prompt

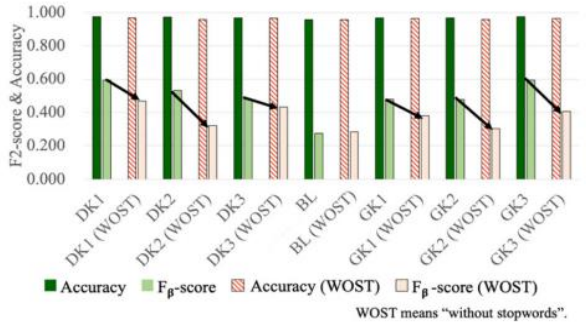


FIGURE 7 Defect detection performance change after the removal of stopwords in a prompt

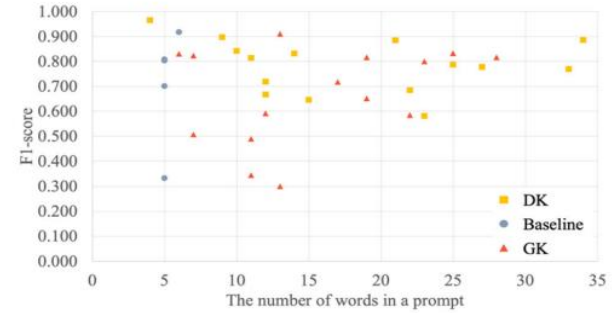


FIGURE 8 Relationship between the defect classification performance (F1-score) and the number of words in a prompt

RESULTS AND ANALYSIS

Visual prompts versus textual prompts

- CLIP을 통해 embedding한 같은 class의 image feature vector를 추가할수록 분류 성능 향상 & 4개 이상부터는 시각적 정보(image)가 더 유용함

- 결함 탐지 또한 CLIP을 통해 embedding한 같은 class의 image feature vector를 추가할수록 탐지 성능 향상 & 5개 이상부터는 시각적 정보(image)가 더 유용함

- 즉, textual promp는 단일 시각적 정보보다 적합한 프롬프트이지만 여러 시각적 정보를 얻을 수 있다면 visual prompt 가 더 적합한 프롬프트

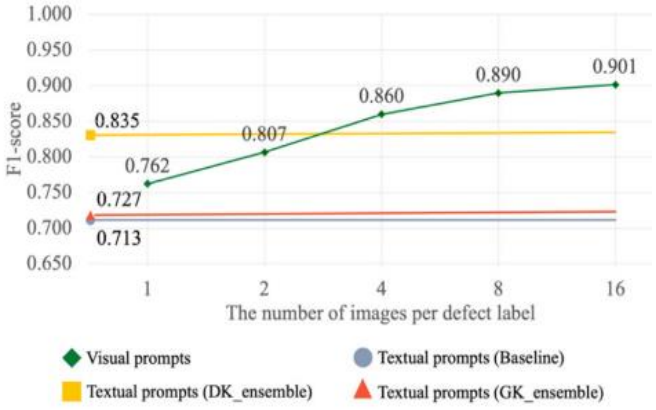


FIGURE 9 Comparison of the defect classification performances between textual and visual prompts

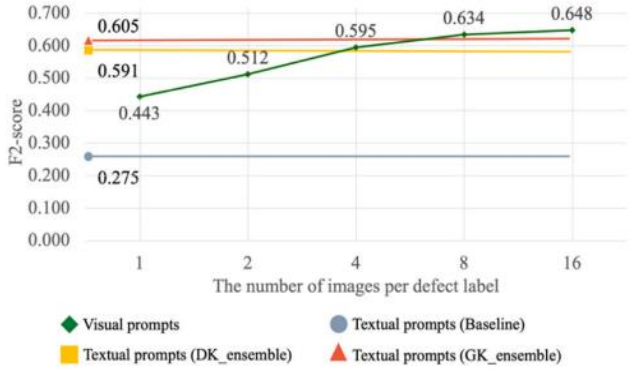


FIGURE 10 Comparison of the defect detection performances between textual and visual prompts

RESULTS AND ANALYSIS

Single-modal prompts versus multimodal prompts

- 시각화 결과, Text와 visual 정보를 embedding 하는 방법에는 차이가 존재
- 분류와 탐지 모두 multi modal이 더 좋은 성능을 보임
- 하지만 앞선 연구와 다르게 GK_ensemble & image 조합이 더 좋은 성능을 보이고 있음
- 앞선 가설과 다른 결과 이지만 Text 와 Visual embedding 정보가 서로 다르기에 이를 보완하며 좋은 성능을 보였다고 판단 가능

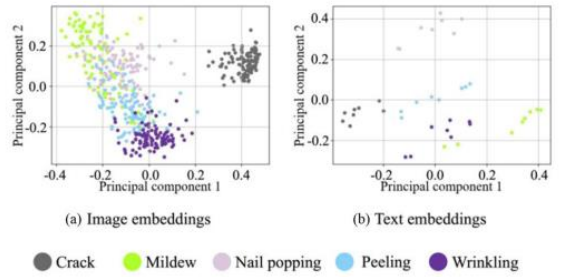


FIGURE 11 Principal component analysis (PCA) results using image and text embedding vectors

TABLE 5 Defect classification performances in ensembling visual and textual prompts

Prompt	Accuracy	F1-score
DK_ensemble (text)	0.834	0.835
GK_ensemble (text)	0.736	0.727
1-shot (image)	0.770	0.762
DK_ensemble + 1-shot (text + image)	0.825	0.821
GK_ensemble + 1-shot (text + image)	0.837	0.834
2-shot (image)	0.817	0.807
DK_ensemble + 2-shot (text + image)	0.864	0.861
GK_ensemble + 2-shot (text + image)	0.870	0.868
4-shot (image)	0.864	0.860
DK_ensemble + 4-shot (text + image)	0.896	0.894
GK_ensemble + 4-shot (text + image)	0.904	0.904
8-shot (image)	0.891	0.890
DK_ensemble + 8-shot (text + image)	0.917	0.917
GK_ensemble + 8-shot (text + image)	0.927	0.928
16-shot (image)	0.902	0.901
DK_ensemble + 16-shot (text + image)	0.925	0.925
GK_ensemble + 16-shot (text + image)	0.933	0.934

TABLE 6 Defect detection performances in ensembling visual and textual prompts

Prompt	Accuracy	F2-score
DK_ensemble (text)	0.975	0.591
GK_ensemble (text)	0.976	0.605
1-shot (image)	0.966	0.433
DK_ensemble + 1-shot (text + image)	0.974	0.570
GK_ensemble + 1-shot (text + image)	0.974	0.567
2-shot (image)	0.970	0.512
DK_ensemble + 2-shot (text + image)	0.976	0.610
GK_ensemble + 2-shot (text + image)	0.976	0.611
4-shot (image)	0.975	0.595
DK_ensemble + 4-shot (text + image)	0.979	0.656
GK_ensemble + 4-shot (text + image)	0.979	0.659
8-shot (image)	0.978	0.634
DK_ensemble + 8-shot (text + image)	0.980	0.674
GK_ensemble + 8-shot (text + image)	0.980	0.671
16-shot (image)	0.979	0.648
DK_ensemble + 16-shot (text + image)	0.981	0.679
GK_ensemble + 16-shot (text + image)	0.981	0.683

Conclusion

DK와 GK를 기반으로 한 프롬프트를 사용해 CLIP을 활용해 zero-shot 학습을 한 결과

- 다음과 같은 contribution을 가짐
  1. DK 가 GK 보다 프롬프트로 더 나은 성능을 발휘
  2. 프롬프트는 핵심 용어 모음보다 완전한 문장 형태가 좋은 성능을 보임
  3. 시각적 정보가 문맥적 정보보다 더 좋은 성능을 보임(단, 단일 시각 정보일 경우 문맥 정보가 더 도움이 됨)
  4. Multi-modal prompt가 single-modal prompt 보다 좋은 성능을 보임

- VLP 모델이 건설 현장에서 분류 혹은 검출 하기 위한 zero-shot & few-shot learning의 모델로 사용 가능성 보임

- 하지만, 여전히 전문가 수준의 세부 결함 정보(결함 심각도, 균열 정도 등)를 식별하는 것은 불가

- 추후 연구를 통해 더 좋은 성능의 assistant AI를 개발할 수 있을 것으로 기대

출처

1. Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image-based structural damage recognition. Computer-Aided Civil and Infrastructure Engineering, 33(9), 748-768.
2. Liang, X. (2019). Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. Computer-Aided Civil and Infrastructure Engineering, 34(5), 415-430.
3. Jiang, S., & Zhang, J. (2020). Real-time crack assessment using deep neural networks with wall-climbing unmanned aerial system. Computer-Aided Civil and Infrastructure Engineering, 35(6), 549-564.
4. Zhu, J., Zhang, C., Qi, H., & Lu, Z. (2020). Vision-based defects detection for bridges using transfer learning and convolutional neural networks. Structure and Infrastructure Engineering, 16(7), 1037-1049.
5. Guo, J., Wang, Q., & Li, Y. (2021). Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification. Computer-Aided Civil and Infrastructure Engineering, 36(3), 302-317.
6. Maeda, H., Kashiyama, T., Sekimoto, Y., Seto, T., & Omata, H. (2021). Generative adversarial network for road damage detection. Computer-Aided Civil and Infrastructure Engineering, 36(1), 47-60.