학습 완료된 분류 모델과 SAM2를 결합한 효율적인 라벨링 방법

송창우, 고영준*
충남대학교
{202460116, yjkoh* }@o.cnu.ac.kr

Efficient Labeling Method Combining Trained Classification Model and SAM2

Chang Woo Song, Yeong Jun Koh*
Chungnam Univ.

요 약

본 논문은 학습 완료된 이미지 분류(Classification) 모델과 Segment Anything Model (SAM) 2 모델을 결합하여, 픽셀(pixel) 단위의 데이터 라벨링(labeling)을 위한 효율적인 방법을 제안한다. 제안하는 방법은 분류 모델에서 생성된 CAM(Class Activation Map)을 분석하여 주요 활성화 영역을 식별하고, 이를 기반으로 SAM2 모델에 입력하여 정밀한 픽셀 단위 라벨(label)을 획득한다. 이러한 방법은 수작업 라벨링의 부담을 줄이고 보다 정확하고 일관된 라벨링을 가능하게 한다. 이를 통하여 이미지 분할(Segmentation) 모델의 성능을 해치지 않으면서, 데이터 라벨링에 소요되는 노력 대비 시간을 줄일 수 있을 것으로 기대한다.

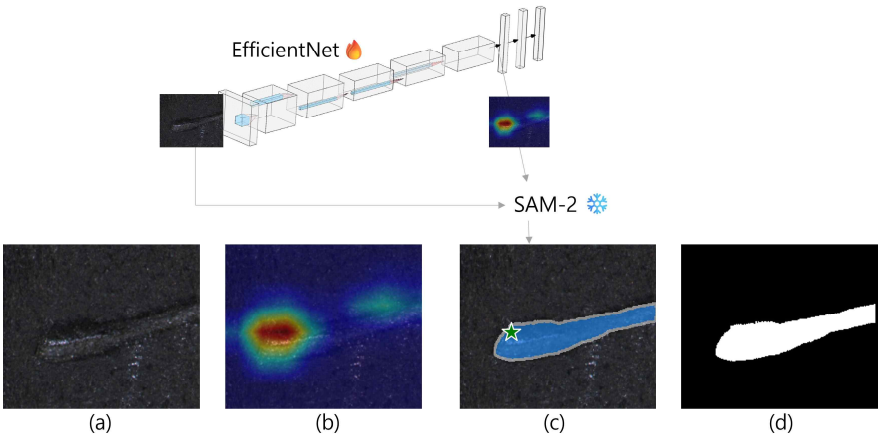

그림1. 제안하는 방법

Ⅰ. 서 론

최근 딥러닝(Deep-learning) 기술의 발전으로, 텍스트 대비 상대적으로 많은 컴퓨팅 자원을 필요로 하는 이미지 기반 컴퓨터 비전(Vision) 과제에서도 눈부신 성과를 이뤄내고 있다. 이미지 분류 문제로 유명한 ImageNet 데이터셋으로 학습한 네트워크들의 성능은 어느덧 사람의 분류 성능보다 높은 결과를 보인다. 이러한 이미지 분류 모델의 성능이 향상됨에 따라, 모델의 결정 과정을 이해하고 해석하려는 노력도 함께 이루어졌다. CAM(Class Activation Mapping)[1]과 같은 기술은 분류 모델이 이미지의 어느 부분에 주목하여 결정을 내리는지 시각화할 수 있게 한다. 이는 모델의 판단 근거를 이해하고 데이터 라벨링 정책을 재 수립하는 등 중요한 도구로 활용될 수 있다. 이미지 분할(Segmentation)은 픽셀 수준에서 이미지의 각 부분을 의미 있는 세그먼트(Segment) 단위의 객체로 구분하는 작업으로, 컴퓨터 비전 분야에서 중요한 과제로 이미지 분류와 분할 기술은 각각 발전하면서도 서로 영향을 주고받았다. 예를 들어, FCN(Fully Convolutional Networks)[2]은 분류 모델에서 사용된 백본(backbone) 네트워크 구조가 분할 모델의 인코더로 활용되었다. 두 분야의 방법들은 상호 보완적으로 발전하고 있다. 또한, 이러한 방법들은 의료 영상 분석, 자율주행 등 다양한 산업 응용 분야에서 중요한 역할을 하고 있다.

최근 등장한 파운데이션 모델(Foundation Model)[3]이라는 새로운 패러다임은 컴퓨터 비전 분야에 큰 변화를 가져오고 있다. 파운데이션 모델은 대규모 데이터셋으로 사전 학습된 거대한 모델로, 다양한 하위 작업에 적용될 수 있는 범용적인 표현을 학습한다. CLIP[4] 과 같은 모델들은 이미지와 텍스트를 함께 학습함으로써, 이미지 분류나 분할과 같은 전통적인 비전 과제를 비롯한 이미지 생성, 이미지와 텍스트간 매칭 등 다양한 작업을 단일 모델로 수행할 수 있다. 이러한 파운데이션 모델의 등장은 컴퓨터 비전 분야의 패러다임을 크게 변화시켜 기존의 task specific한 접근 방식에서 벗어나, 하나의 거대 모델이 다양한 하위 작업에 적용될 수 있는 가능성을 보여주고 있다. 이는 이미지 분류, 객체 탐지, 이미지 분할 등 다양한 비전 작업들이 더욱 통합적으로 발전할 수 있는 기반을 마련하고 있다.

본 논문에서는 학습 완료된 이미지 분류 모델과 파운데이션 모델인 SAM2[5]를 결합한 효율적인 데이터 라벨링 방법을 제안한다. 제안하는 방법은 분류 모델에서 생성된 CAM을 활용하여 각 클래스의 주요 활성화 영역을 식별한 뒤, 이를 SAM2 모델에 입력으로 활용하여 정밀한 픽셀 단위 라벨을 획득한다.

Ⅱ. 제안하는 방법

그림 1은 본 논문에서 제안하는 방법을 전체적으로 나타낸다. 학습이 완료된 분류 모델에서 그림 1. (b)와 같이 CAM을 획득하여, 파운데이션 모델인 SAM-2의 입력으로 활용한다. 그 결과는 그림 1. (c)와 같다. 그리고 획득된 결과를 이진화(thresholding)하여 그림 1. (d)와 같이 해당되는 세그먼트가 라벨링 된 마스크 이미지를 얻는다.

본 논문에서는 CAM이후 제안된 다양한 방법들[6][7] 중 Grad-CAM[6]을 활용한다. 수식 1은 Grad-CAM 계산 과정과 이후 SAM2의 프롬프트로 활용할 포인트(point)를 구하는 의사코드(pseudo code)를 나타낸다.

수식 1. 프롬프트 포인트 획득 과정

Input: Feature maps A^k, class c
Output: Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$

1: Compute importance weights:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

2: Generate Grad-CAM:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

3: Identify the maximum activation point:

$$(i^*, j^*) = \arg\max_{i,j}\left(L^c_{\text{Grad-CAM}}(i,j)\right)$$

Ⅲ. 실험 결과

객관적인 성능 평가를 위해, 산업 현장과 유사한 환경에서 획득된 표면 결함에 관한 KSSD2[8] 데이터셋을 활용한다. 데이터셋에 포함된 이미지의 개수는 결함이 있는 356 장과 결함이 없는 2,979 장으로 구분된다. 결함이 있는 이미지는 같은 해상도로 이진화 된 결함이 표시된 마스크 이미지가 포함되어 있다. 본 논문에서는 해당 데이터셋을 이미지 분류 모델인 EfficientNet[9]으로 학습하고, 결함이 있는 이미지에 대해 제안하는 방법으로 획득한 마스크 이미지를 분할 네트워크의 입력으로 활용한다.

표 1은 결함이 있는 영상 1장을 대상으로 일관성 있는 픽셀 단위의 데이터 라벨링을 수행할 때 데이터셋 별 최소로 필요한 시간을 나타낸다. 각 이미지 별 1분이 소요된다고 가정했을 때, 최소 필요 시간은 131분 정도 차이가 발생함을 확인할 수 있다.

표 1. 데이터 라벨링 최소 필요 시간 (단위: 분)

	Train	Test
manual	246	110
proposed	156	69

참 고 문 헌

[1] Zhou, B., et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[2] Long, J., et al. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[3] Bommasani, R., et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

[4] Radford, A., et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[5] Ravi, N., et al. "Sam 2: Segment anything in images and videos." arXiv preprint arXiv:2408.00714 (2024).

[6] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[7] Wang, H., et al. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

[8] Bož ič , Jakob, Domen Tabernik, and Danijel Skoč aj. "Mixed supervision for surface-defect detection: From weakly to fully supervised learning." Computers in Industry 129 (2021): 103459.

[9] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.