클래스분류 학습이 Self-Supervised Transformer의 saliency map에 미치는 영향 분석

김재욱[O,*], 김현철[**]
[O]고려대학교 인공지능융합학과,
[*]에이모 AI Lab,
[**]고려대학교 컴퓨터학과
e-mail: lesit.jae@gmail.com[O,*], harrykim@korea.ac.kr[**]

Analysis of the effect of class classification learning on the saliency map of Self-Supervised Transformer

JaeWook Kim[O,*], Hyeoncheol Kim[**]
[O]Dept. of Applied Artificial Intelligence, Korea University,
[*]AI Lab, AIMMO,
[**]Dept. of Computer Science and Engineering, Korea University

● 요 약 ●

NLP 분야에서 적극 활용되기 시작한 Transformer 모델을 Vision 분야에서 적용하기 시작하면서 object detection과 segmentation 등 각종 분야에서 기존 CNN 기반 모델의 정체된 성능을 극복하며 향상되고 있다. 또한, label 데이터 없이 이미지들로만 자기지도학습을 한 ViT(Vision Transformer) 모델을 통해 이미지에 포함된 여러 중요한 객체의 영역을 검출하는 saliency map을 추출할 수 있게 되었으며, 이로 인해 ViT의 자기지도학습을 통한 object detection과 semantic segmentation 연구가 활발히 진행되고 있다. 본 논문에서는 ViT 모델 뒤에 classifier를 붙인 모델에 일반 학습한 모델과 자기지도학습의 pretrained weight을 사용해서 전이학습한 모델의 시각화를 통해 각 saliency map들을 비교 분석하였다. 이를 통해, 클래스 분류 학습 기반 전이학습이 transformer의 saliency map에 미치는 영향을 확인할 수 있었다.

키워드: Transformer, Self-Supervised Learning, Transfer Learning, Saliency map

I. Introduction

논문 'Emerging Properties in Self-Supervised Vision Transformers'[2]에서 teacher & student 기반 자기지도학습에 새로 고안한 DINO라는 알고리즘을 적용하여 top-1 acc가 80.1%에 달하는 최고의 분류 성능을 달성하였다.

본 논문에서는 자기지도학습한 DINO pretrained weight을 기반으로 전이학습하여 테스트 데이터를 multi class classification(하나의 class만 선택하는 분류)과 multi label classification(여러 class 선택이 가능한 분류)으로 학습한 후에, 테스트 데이터로 추론(inference) 후 각각 object에 활성화하는 saliency map을 추출하여 semantic segmentation검출 사용 가능성을 확인하기 위해 시험을 진행하였다. 이를 위해 각 classifier를 ViT의 뒤에 붙여 ViT의 마지막 블록에 있는 self-attention을 입력으로 하는 구조를 만들어 지도학습을 할 수 있도록 했다. 또한, DINO를 통해 자기지도학습된 pretrained weight을 사용한 전이학습과 일반지도학습, multi class classification 학습과 multi label classification 학습 등 여러 학습 방법에 따라 추출된 saliency map을 비교함으로써 성능향상 가능성을 분석하였다. 이를 통해, classification 학습이 객체의 saliency map 검출에 방해되는 문제와 원인을 유추할 수 있었다.

II. Preliminaries

1. Related works

1.1 Vision Transformer

Transformer는 NLP분야에서 제안되어 많이 사용되고 있는데, 이를 Vision 분야에 적용함으로서 Clustering, Classification, Object Detection, Segmentation, 3D Object Detection 등 다양한 모델에서

CNN을 대체하며 높은 성능(SOTA)을 내기 시작했다[1].

Transformer는 더 많은 데이터셋을 학습하기 위해 모델의 크기를 확장해도 포화(saturated)되는 문제가 없이 성능이 높아짐과 동시에 리소스(resource) 대비 계산속도 향상의 장점을 가지는데, 이 장점으로 CNN의 정체된 한계를 극복하는 계기가 되었다. 적은 데이터로 새로 학습했을 경우 제한되는 성능 문제로 인해 대량의 데이터로 사전 학습된 weight을 사용해야 한다는 한계점이 있지만, 전이학습이 새로운 데이터의 적용을 빠르게 해주며 더욱더 성능을 높이기 때문에 문제점이라고 볼 순 없다.

1.2 Self-supervised Vision Transformer

여기에서는 논문 'Emerging Properties in Self-Supervised Vision Transformers'[2]에서 제안한 ViT(Vision Transformer)에 NCE(Noise Contrastive Estimator)를 사용한 평가 방법과 teacher & student network의 knowledge distillation 방법, 그리고 여기에서 제안한 DINO라는 알고리즘을 적용하여 자기지도학습을 하도록 하였고, DINO를 적용하지 않았을 때 KNN-classifier의 성능 측정 결과 acc(accuracy)가 78.3% 나왔던 성능이 DINO를 적용했을 때 80.1%로 높아졌다. DINO의 구조는 아래 Fig.1.과 같다.

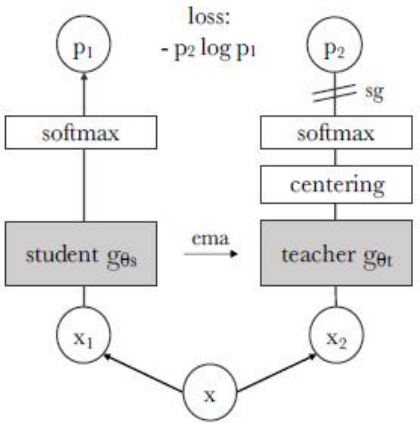

Fig. 1. DINO structure

III. The Proposed Scheme

1. 모델 구조

1.1 ViT 구조의 크기

ViT는 크기에 따라 base, small, tiny로 구분되는데 여기에서는 base 구조에 비해 성능이 큰 차이가 없으면서 학습 속도가 빠른 small 구조를 사용했다. 각 영역의 특징을 추출하기 위한 patch(window) 크기는 성능을 높이기 위해 8을 사용하였다.

1.2 Classifier 추가

본 논문에서는 DINO에서 사용한 eval linear의 방식을 사용했으며, ViT의 마지막 블록에 있는 elf-attention을 입력으로 하는 classifier를 붙였다. loss function은 multi class classification의 경우 cross entropy함수를 사용하며 multi label classification의 경우엔 각 class 별로 binary classification을 붙이는 BCE(binary cross entropy)를 사용했다.

1.3 Visualization

DINO에서 모델의 마지막 블록에서 self-attention module을 통해 saliency map을 구해서 시각화하였는데, DINO를 기반으로 객체의 localize 방법을 제시한 논문 "Localizing Objects with Self-Supervised Transformers and no Labels"[2]에서는 self-attention module에 있는 linear module인 qkv(Q[query], K[key], V[value]로 구성됨) 중 k feature를 사용하여 이미지의 패치간의 유사도를 계산하여 사용하였다.

본 논문에서는 이 LOST(Localizing Objects with Self-Supervised Transformers and no Labels)의 시각화 방법을 사용하였으며, 이 유사도 계산결과를 사용하여 correlation을 계산하여 saliency map의 영역을 구해 시각화하는 것과 유사도의 degree를 사용하여 활성화의 정도를 구해 시각화하는 것을 사용하였다.

2. 테스트

2.1. Weakly supervised learning

모델을 DINO에서 사전학습한 pretrained weight을 사용하여 테스트 데이터셋으로 전이학습하는 weakly supervised learning 방식으로 실험을 진행하였다.

2.2 Classification learning

차량의 종류만 존재하는 Stanford 대학의 stanford cars 데이터셋을 사용해서 one-hot 방식인 multi class classification 학습을 진행했다.

DINO의 multi classification 뿐만 아니라 multi label classification 학습도 진행했으며, 여러 개의 객체 종류(차와 보행자, 자전거, 오토바이 등)로 구분되어 bbox(bounding box) label이 있는 bdd100k 데이터셋이 하나의 이미지에 여러 개의 객체가 함께 포함되는 것을 착안하여, 각 객체의 class 정보를 사용한 multi-hot 방식인 multi label classification 학습을 진행했다.

2.3 지도학습 실험 추가

추가로 전이학습을 하지 않고 직접 데이터셋으로부터 학습한 모델의 추론 결과도 함께 비교하기 위해, DINO를 bdd100k 데이터셋으로만 자기지도학습한 모델로 saliency map을 시각화하였다.

2.4 학습 및 테스트 데이터

각 학습에는 stanford cars 데이터셋과 bdd100k 데이터셋을 사용했으며 visualization을 위한 테스트 데이터는 bdd100k를 사용하여 비교했다.

아래 그림 Fig.2. 는 차량의 종류만 label 되어 있는 stanford

cars 데이터셋의 이미지 예제이고, 그림 Fig3.은 주행영상에 도로의 여러 객체에 대한 bbox와 클래스가 label 되어 있는 bdd100k 데이터셋 이미지 예제이다.

Fig. 2. stanford cars 데이터셋 예

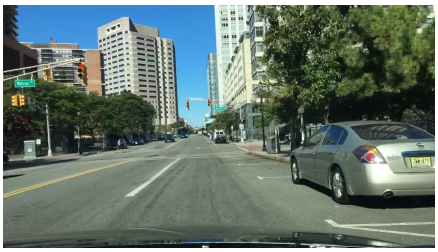

Fig. 3. stanford cars 데이터셋 예

3. 테스트 결과

3.1 DINO pretrained weight, bdd100k train

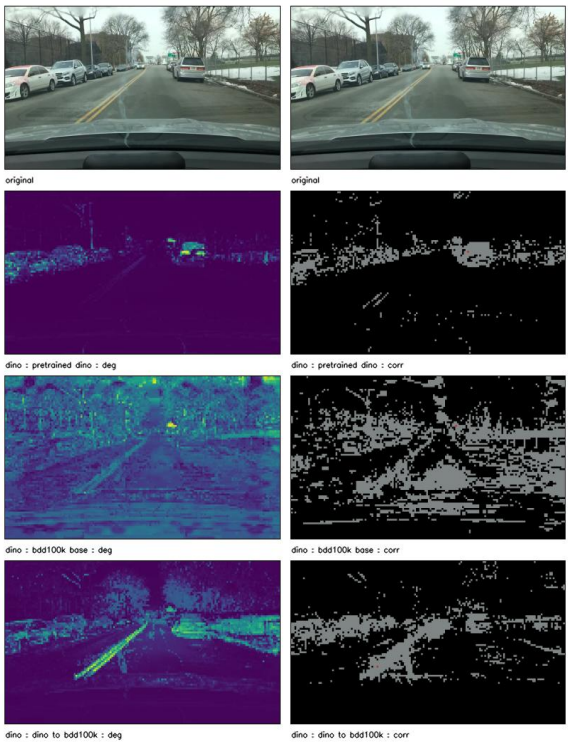

Fig. 4. DINO 학습 방식별 degree, correlation 시각화

위 그림 Fig. 4. 은 degree를 사용한 시각화와 correlation 시각화 그림이며, DINO의 사전학습 모델을 사용한 시각화(dino : pretrained dino)의 표현이 bdd100k로 처음부터 학습한 모델(dino : bdd100k base)을 시각화한 것보다 차량들에 더 활성화 된 것을 볼 수 있다. DINO pretrained weight을 사용하여 전이학습한 모델(dino : dino to bdd100k)은 degree 시각화 상으로는 더 활성화 되었지만, 차량뿐만 아니라 bdd100k 클래스에 포함되어 있지 않는 나무와 차선 등 다른 부분에도 활성화되면서, 오히려 덜 중요한 부분까지 saliency map에 포함되었다.

3.2 multi class transfer learning: stanford cars

그림 Fig.5. 는 stanford cars 데이터셋으로 dino pretrained weights을 전이학습한 모델의 시각화 결과와 비교한 그림이다. dino pretrained weight의 결과가 degree 시각화에서 차량에 더 잘 활성화 되고 있고, correlation 시각화로는 약간의 차이로 차량을 검출하고 있다. 참고로, 그림 중 맨 아래는 correlation결과로 영역을 추출하여 이 영역에만 degree 시각화 결과를 채운 결과이다.

3.3 multi label class transfer learning: bdd100k

그림 Fig.6. 은 bdd100k으로 dino pretrained weight을 전이학습한 모델의 시각화 결과와 비교한 그림이다. bdd100k로 학습한 모델의 시각화 결과가 stanford cars로 학습한 모델의 시각화 결과보다 차량에 덜 활성화되는 것을 볼 수 있다.

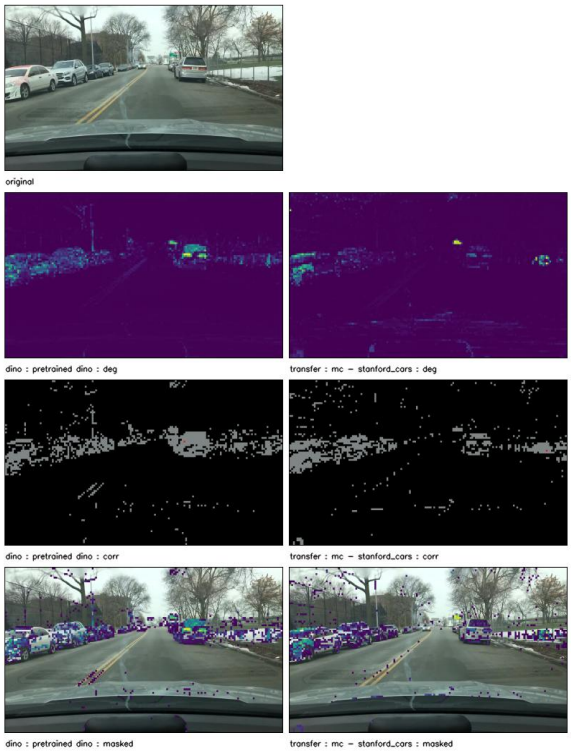

Fig. 5. stanford cars 전이학습 correlation 시각화

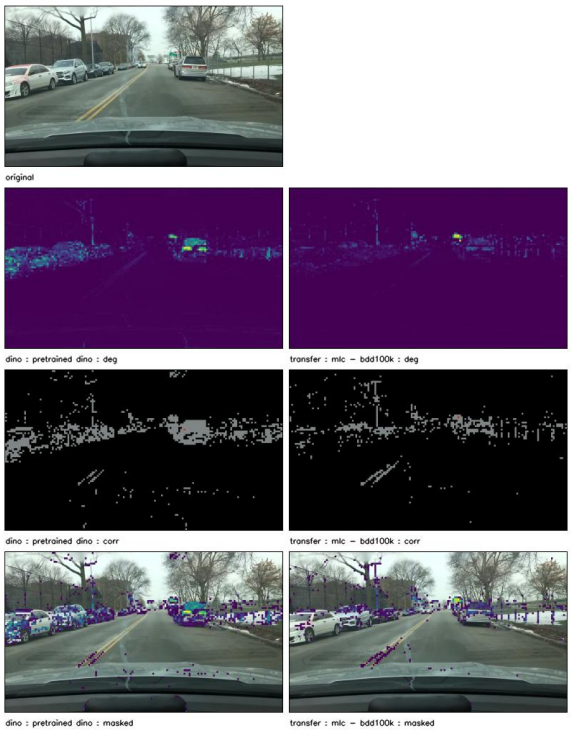

Fig. 6. bdd100k 전이학습 correlation 시각화

결과적으로 label이 없이 DINO로 자기지도학습한 Vision Transformer 모델이 객체(주로 차량)에 더 활성화하는 saliency map을 추출하였으며, 다양한 종류의 class로 label된 주행영상의 데이터셋인 bdd100k 로 multi label classification 기반 전이학습했을 경우 label된 class의 객체에 대한 활성화는 낮아지고 오히려 label 되어 있지 않는 차선과 나무 같은 배경의 요소에도 활성화되는 것을 볼 수 있었다. 이런 원인은 하나의 이미지에 차량과 보행자 등이 하나씩만 존재하지 않고 여러 객체가 동시에 나타나기 때문에 multi label classification으로는 label의 클래스와 객체를 매칭하지 못하는 문제가 발생하면서 활성화 정도가 낮아지고, 모든 영상에 차선과 나무가 있기 때문에 이런 배경의 요소까지 활성화되기 때문인 것으로 유추할 수 있다.

IV. Conclusions

본 연구를 통해, 전이학습을 하였음에도 사용한 데이터셋의 label이 클래스로만 이미지를 설명함으로서 classification 학습이 label외의 배경 요소까지 활성화 시키는 영향을 확인할 수 있었다. 이를 통해 bbox label을 사용했을 경우 원하는 객체의 localize 정보를 가진 bbox label을 사용했을 경우 객체에 대한 높은 활성화와 정교한 검출 영역을 기대해 볼 수 있으며, 의미 있는 결과가 나올 경우 비싼 비용을 가지는 segmentation label 작업을 대체하는 연구도 기대해 볼 수 있다. 추후 연구에서 ViT뒤에 classifier가 아닌 object detection을 수행하는 detector를 붙여서 원하는 객체에 대한 localize 정보 추출의 가능성을 연구할 계획이다.

REFERENCES

[1] Alexey Dosovitskiy et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“, Proceedings of the ICLR Conference, 2021.

[2] Mathilde Caron et al, "Emerging Properties in Self-Supervised Vision Transformers", Apr 2021.

[3] Oriane Siméoni et al, "Localizing Objects with Self-Supervised Transformers and no Labels", Journal of BMVC, pp3~4, Sep. 2021.