💚

[16주차] Vision Transformers Need Registers

1. Introduction

- 문제 배경
  - 이미지 임베딩의 중요성
    - 이미지에서 추출한 특징을 다양한 컴퓨터 비전 작업에 사용할 수 있게 하는 것은 오랜 문제였음
    - 초기 방법들은 SIFT와 같은 수작업 원칙에 의존함

> 🌟 **SIFT(Social Information Forensics Toolkit)**
>
> 일반적으로 소셜 미디어에서의 정보 검증을 위한 도구와 방법론을 제공하는 프레임워크
> 1️⃣
> **Stop(멈추기)**: 섣불리 행동 ❌
> 2️⃣
> **Investigate the Source(출처 조사하기)**: 정보를 제공한 출처를 조사
> 3️⃣
> **Find Better Coverage(더 나은 취재 찾기)**: 동일한 정보에 대해 다른 신뢰할만한 출처들이 어떻게 보도하고 있는지 찾기
> 4️⃣
> **Trace Claims, Quotes, and Media to the Original Context (주장, 인용, 미디어의 원래 맥락 추적하기)**: 정보의 출처가 인용한 주장, 인용문, 이미지, 비디오 등이 원래 어떤 맥락에서 사용되었는지를 추적하기

  - 데이터와 딥러닝의 발전

        - 대규모 데이터와 딥러닝 기술의 발전으로 인해 이제는 엔드 투 엔드 학습이 가능해짐
    - 일반적인 특징 임베딩의 필요성
        - 특정 작업을 위한 라벨링된 데이터 수집이 어렵기 때문에, 일반적인 특징 임베딩은 여전히 중요함
        - 예를 들어 의료 데이터나 원격 탐사 데이터는 전문 지식이 필요하거나 대규모로 수집하는 데 비용이 많이 듦
- 사전 학습과 특징 추출
    - 사전 학습 모델: 오늘날에는 많은 데이터가 있는 작업에 대해 모델을 사전 학습시킨 후, 이 모델의 일부를 특징 추출기로 사용하는 것이 일반적임
    - 지도 학습 방법: 분류나 텍스트-이미지 정렬을 통해 강력한 특징 모델을 학습시켜, 후속 작업을 수행할 수 있음
    - **트랜스포머 아키텍처**를 기반으로 하는 **자가 지도 학습 방법**은 후속 작업에서 높은 예측 성능을 보여 주목받고 있음
- DINO 알고리즘
    - DINO 알고리즘: 이미지의 의미적 레이아웃에 대한 명시적인 정보를 포함하는 모델을 생성하는 것으로 나타남
    - **attention map**: 마지막 attention 레이어는 이미지의 의미적으로 일관된 부분에 자연스럽게 집중하며, 해석 가능한 attention map을 자주 생성함
    - DINO를 기반으로 한 LOST와 같은 알고리즘은 attention map의 정보를 모아 객체를 비지도 방식으로 탐지할 수 있음

DINO(Self Distillation with No Labels)

- 자가 지도 학습(Self-Supervised Learning) 방법 중 하나로, 지도 학습 없이 데이터에서 유용한 표현을 학습하는 것을 목표로 함
- "Self-Distillation with No Labels"라는 이름에서 알 수 있듯이, 라벨이 없는 데이터로 학습함

1️⃣
자가 증류(Self Distillation): 교사 모델과 학생 모델이라는 두 개의 신경망을 사용함. 교사 모델은 고정된 파라미터를 가지고 있으며, 학생 모델은 교사 모델을 모방하도록 학습됩니다. 이 과정에서 학생 모델은 교사 모델의 출력을 최대한 닮도록 학습됨

2️⃣
멀티 크기 입력: 다양한 크기의 입력 이미지를 사용하여 모델을 학습시키며, 이는 모델이 다양한 스케일에서 일관된 표현을 학습하도록 도움

3️⃣
비지도 학습(No Labels): 라벨이 없는 데이터셋을 사용하여 학습하며, 이는 많은 양의 라벨링된 데이터가 필요 없는 장점이 있음

4️⃣
비교 학습(Contrastive Learning): 두 개의 서로 다른 증강된 버전의 이미지를 비교하여 좋은 표현을 학습하며, 이는 모델이 이미지의 중요한 특징을 학습하도록 도움

LOST(Localization without Supervision)

- 비지도 학습을 통해 이미지 내 객체를 탐지하는 알고리즘
- 지도 학습 없이도 이미지에서 의미 있는 객체를 찾을 수 있는 방법을 제안함

1️⃣
그래프 기반 접근(Grapical Approach): 이미지를 그래프로 표현하며, 여기서 각 노드는 이미지의 패치를 나타내고 간선은 패치 간의 유사성을 나타냄
2️⃣
노드 중심성: 그래프에서 각 노드의 중심성을 계산하여 중요한 노드를 찾으며, 높은 중심성을 가진 노드는 이미지 내에서 중심적이고 중요한 위치를 차지할 가능성이 큼
3️⃣
이미지 패치(Image Patch): 이미지를 작은 패치로 나누고, 각 패치의 특징을 추출하여 그래프를 구성함. 이를 통해 이미지 내에서 중요한 객체를 찾음

- DINOv2와 문제점
  - DINOv2: DINO의 후속 버전으로, 밀집 예측 작업을 처리할 수 있는 특징을 제공함
  - DINOv2는 LOST와 호환되지 않는 문제를 보였으며, 특징을 추출할 때 DINOv2는 실망스러운 성능을 보였고, 이는 DINO와는 다른 동작을 한다는 것을 시사함

- 아티팩트의 발견과 분석
  - DINOv2의 특징 맵에서 첫 번째 버전에서는 나타나지 않았던 아티팩트가 존재함을 발견했음
  - 이 아티팩트는 높은 노름을 가진 토큰으로, 시퀀스의 약 2%를 차지함. 이러한 토큰은 트랜스포머의 중간 레이어에서 나타나며, 충분히 큰 트랜스포머가 충분히 오래 학습된 후에 나타남
  - 아티팩트 토큰은 원래 위치나 패치의 원본 픽셀에 대한 정보를 덜 포함하고 있으며, 이는 모델이 추론 중에 이러한 패치의 지역 정보를 버린다는 것을 시사함

> 🌟 **아티팩트(artifact)**
>
> - 주로 데이터 처리나 모델 학습 과정에서 의도하지 않게 발생하는 이상 현상이나 오류를 의미함
> -
> **DINOv2 알고리즘**에서 발견된 아티팩트: 높은 노름을 가진 토크(다른 토큰들에 비해 비정상적으로 큰 값), 시퀀스의 약 2%를 차지, 트랜스포머의 중간 레이어에서 발생하며 모델이 크고 충분히 오래 학습된 후에 나타난다는 특징을 지님

- 아티팩트 해결 방법
  - 레지스터 토큰 추가: 입력 이미지와 독립적으로 **토큰 시퀀스에 레지스터 토큰을 추가**하여 이 문제를 해결할 수 있음
  - 이러한 수정을 거친 후 아티팩트 토큰이 시퀀스에서 완전히 사라졌으며, 모델의 밀집 예측 작업에서 성능이 향상되고 특징 맵이 더욱 부드러워짐

2. Problem Formulation

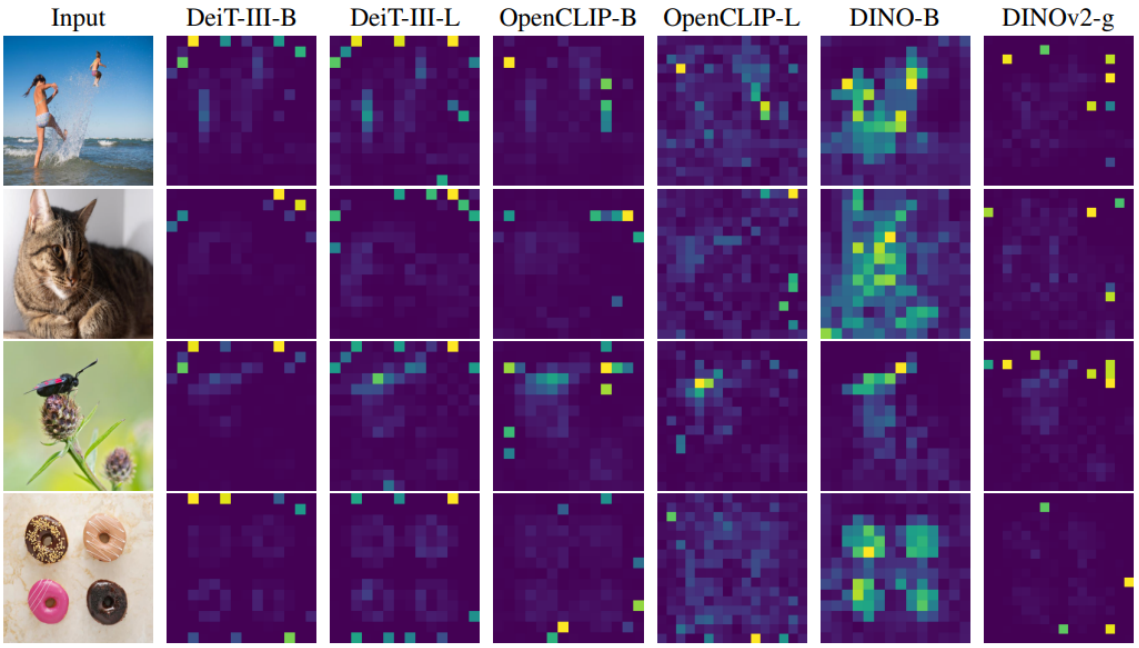


Figure 2: Illustration of artifacts observed in the attention maps of modern vision transformers. We consider ViTs trained with label supervision (DeiT-III), text-supervision (OpenCLIP) or self-supervision (DINO and DINOv2). Interestingly, all models but DINO exhibit peaky outlier values in the attention maps. The goal of this work is to understand and mitigate this phenomenon.

- 위 Fig2에서 볼 수 있듯, 대부분의 최신 비전 트랜스포머들은 attention map에서 아티팩트를 보임
- 비지도 학습 기반인 DINO 백본 모델은 local features(세부적 특징)의 품질과 attention map의 해석 가능성 측면에서 높은 평가를 받음
- 그러나 후속 모델인 DINOv2는 좋은 로컬 정보를 유지하고 있음에도 불구하고 attention map에서 원치 않는 아티팩트가 나타나는 것으로 밝혀짐

➡️ **이러한 아티팩트가 왜, 언제 나타나는지를 연구하는 것이 목적**

2.1 Artifacts in the local features of DINOv2

Artifacts are high-norm outlier tokens

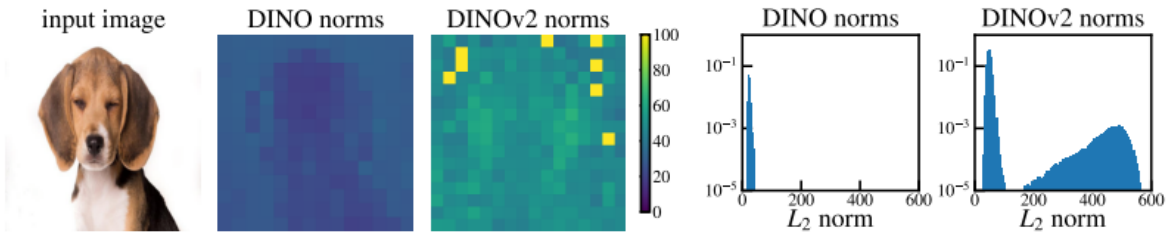
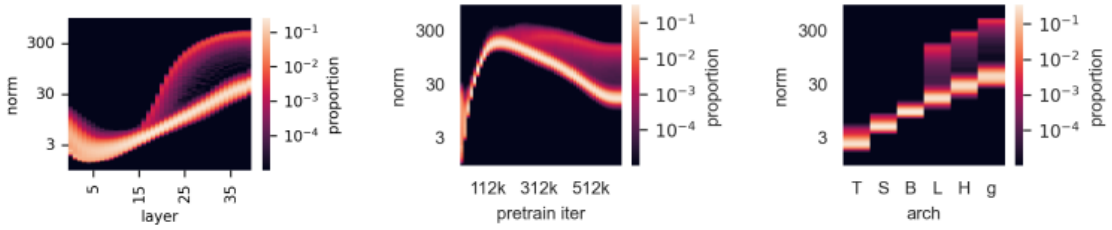


Figure 3: Comparison of local feature norms for DINO ViT-B/16 and DINOv2 ViT-g/14. We observe that DINOv2 has a few outlier patches, whereas DINO does not present these artifacts. For DINOv2, although most patch tokens have a norm between 0 and 100, a small proportion of tokens have a very high norm. We measure the proportion of tokens with norm larger than 150 at 2.37%.

- 로컬 특징에서 나타나는 아티팩트를 정량적으로 특성화하고자 함
- 아티팩트 패치와 다른 패치의 중요한 차이는 **모델 출력에서 토큰 임베딩의 노름**(모델이 출력하는 벡터의 크기)
- Fig. 3 (왼쪽)에서 DINO와 DINOv2 모델의 로컬 특징 노름을 비교한 결과, DINOv2에서 아티팩트 패치의 노름이 다른 패치들보다 훨씬 높다는 것을 알 수 있음
- Fig. 3 (오른쪽)에서는 작은 데이터셋의 이미지들에 대한 특징 노름의 분포를 나타내고 있으며, 이 분포는 명확히 이중봉형(bimodal)
    - 이중봉형 분포: 두 개의 뚜렷한 봉우리를 가진 분포로, 두 개의 다른 그룹이 존재함을 시사함
- 노름이 150을 넘는 토큰들을 "**high-norm**" 토큰으로 간주하고, 이들의 특성을 일반 토큰들과 비교하여 연구함
    - 본 논문에서 high-norm = 이상치

Outliers appear during the training of large models



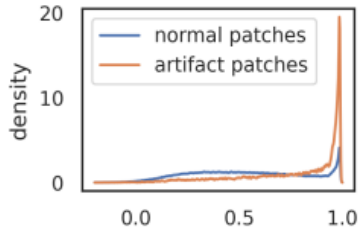
(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

Figure 4: Illustration of several properties of outlier tokens in the 40-layer DINOv2 ViT-g model. **(a)**: Distribution of output token norms along layers. **(b)**: Distribution of norms along training iterations. **(c)**: Distribution of norms for different model sizes. The outliers appear around the middle of the model during training; they appear with models larger than and including ViT-Large.

- DINOv2 모델 학습 중 이상치 패치가 나타나는 조건에 대한 추가 관찰

- 40개 레이어 중 **15번째 레이어**에서 high-norm 패치가 다른 패치들과 차별화되기 시작함
- DINOv2 학습 과정에서 high-norm 패치가 **학습의 1/3 지점 이후**에 나타남
- Tiny, Small, Base, Large, Huge, Giant 크기의 모델을 분석한 결과, **가장 큰 세 모델**에서만 이상치가 나타남

High-norm tokens appear where patch information is redundant



(a) Cosine similarity to neighbors.

	position prediction		reconstruction
	top-1 acc	avg. distance ↓	L2 error ↓
normal	**41.7**	**0.79**	**18.38**
outlier	22.8	5.09	25.23

(b) Linear probing for local information.

Figure 5: **(a)**: Distribution of cosine similarity between input patches and their 4 neighbors. We plot separately artifact patches (norm of the *output token* over 150) and normal patches. **(b)**: Local information probing on normal and outlier patch tokens. We train two models: one for predicting position, and one for reconstructing the input patch. Outlier tokens have much lower scores than the other tokens, suggesting they are storing less local patch information.

- 패치 임베딩 바로 후에 연구자들이 high-norm 토큰과 그 이웃 패치들 사이의 코사인 유사도를 측정
- 그 결과, high-norm 토큰이 이웃 패치들고 매우 유사한 곳에서 발생한다는 것을 보임

➡️ **high-norm이 중복된 정보를 포함하고 있으며, 모델이 이러한 토큰의 정보를 버려도 이미지 표현의 품질에 큰 영향을 미치지 않는다는 것을 의미함**

High-norm tokens hold little local information

- 연구자들은 high-norm 토큰에 대해 더 잘 이해하기 위해 패치 임베딩을 다양한 정보 유형에 대해 탐색했음
  - 위치 예측: 선형 모델을 훈련시켜 각 패치 토큰의 위치를 예측하고 정확도를 측정함. high-norm 토큰은 다른 토큰보다 **훨씬 낮은 정확도**를 보여주었으며 이는 high-norm 토큰이 **이미지 내 위치에 대한 정보를 덜 가지고 있음**을 시사함
  - 픽셀 재구성: 선형 모델을 훈련시켜 패치 임베딩으로부터 이미지의 픽셀 값을 예측하고 모델의 정확도를 측정함. high-norm 토큰은 다른 토큰보다 **훨씬 낮은 정확도**

를 보여주었으며 이는 high-norm 토큰이 **이미지 재구성에 필요한 정보를 덜 가지고 있음**을 시사함

Artifacts hold global information

	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	**86.0**	**66.4**	**87.3**	**99.4**	**94.5**	**91.3**	96.9	**91.5**	**85.2**	**99.7**	**94.7**	**96.9**	**78.6**	89.1
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	69.0	55.1	79.1	99.3	93.7	84.9	**97.6**	85.2	84.9	99.6	93.5	94.1	78.5	**89.7**

Table 1: Image classification via linear probing on normal and outlier patch tokens. We also report the accuracy of classifiers learnt on the class token. We see that outlier tokens have a much higher accuracy than regular ones, suggesting they are effectively storing global image information.

- 연구자들은 high-norm 토큰이 얼마나 많은 전역 정보를 가지고 있는지 평가하기 위해 표준 이미지 표현 학습 벤치마크를 사용함
  - 각 이미지에 대해 DINOv2-g를 통해 패치 임베딩을 추출하고, 고노름 토큰이나 일반 토큰 중 임의로 하나를 선택하여 이미지 표현으로 사용
  - 이 표현을 기반으로 로지스틱 회귀 분류기를 훈련시켜 이미지 클래스를 예측하고, 정확도를 측정

➡️ high-norm 토큰이 다른 토큰보다 **훨씬 높은 정확도**를 보여주었으며 이는 high-norm 토큰이 다른 패치 토큰보다 **더 많은 전역 정보를 포함하고 있음**을 시사함

2.2 Hypothesis and Remediation

Hypothesis

- 충분히 훈련된 대형 모델은 중복된 토큰을 인식하고, 이들을 전역 정보를 저장, 처리, 검색하는 데 사용한다는 가설을 세움
- 이러한 행동 자체는 문제가 아니지만, 패치 토큰 안에서 발생하는 것이 문제 → 패치 토큰이 지역적인 정보를 잃게 되어 이미지의 **세부적인 정보를 무시**하게 되고, 이는 **밀집 예측 작업에서 성능 저하를 초래**할 수 있음
  - 밀집 예측 작업: 이미지나 비디오 같은 입력 데이터의 각 픽셀 또는 작은 부분에 대해 예측을 수행하는 작업

Remediation

- 새로운 토큰 추가: 모델이 레지스터로 사용할 수 있는 새로운 토큰을 명시적으로 추가

    - 패치 임베딩 레이어 이후에 추가, 학습 가능한 값을 가짐(CLS 토큰과 유사)
    - 비전 트랜스포머의 마지막에서는 이 새로운 토큰들은 버려지고, [CLS] 토큰과 패치 토큰이 이미지 표현으로 사용됨
- 불확실성 및 추가 연구 필요성
    - 훈련 과정에서 모델에서 나타나는 아티팩트(훈련 중 발생하는 이상 현상)가 어떤 부분에서 발생하는지 완전히 규명하지는 못함
    - 사전훈련 방법이 중요한 역할을 하는 것으로 보이며, OpenCLIP과 DeiT-III 모델에서 이러한 아티팩트가 나타났음
    - 델의 크기와 훈련 길이도 중요한 요소로 작용하는 것을 관찰했음

3. Experiments

- 솔루션 검증 법과 실험의 주요 단계

  1️⃣ 제안된 솔루션 검증 → 레지스터 토큰 도입으로 성능 개선

  2️⃣ 정량적 및 정성적 분석(다양한 지표 + 시각적 결과로 성능 평가)

  3️⃣ 레지스터 개수의 영향 분석 → 최적의 레지스터 토큰 개수 찾기

  4️⃣ 비지도 객체 발견 방법 평가

  5️⃣ 레지스터가 학습한 패턴에 대한 정성적 분석 → 레지스터 토큰이 특정한 이미지 패턴이나 특징을 어떻게 학습했는지 시각적으로 확인

3.1 Training Algorithms and Data

DEIT-III

- ImageNet-1k와 ImageNet-22k 데이터셋에서 ViTs(비전 트랜스포머)를 사용한 간단하고 강력한 지도 학습 레시피
- 레이블이 있는 데이터를 사용한 학습 방법의 예로 선택됨
    - 단순하고 기본 ViT 아키텍처를 사용하며, 강력한 분류 성능을 달성하고 재현 및 개선이 용이함
- ImageNet-22k 데이터셋을 사용하여 ViT-B 설정으로 실험을 수행

OpenCLIP

- 원래 CLIP 작업을 따르며, 텍스트-이미지 정렬 모델을 생성하기 위한 강력한 학습 방법

- 텍스트-감독 학습 방법의 예로 선택됨
    - 오픈소스이며 기본 ViT 아키텍처를 사용하고, 재현 및 개선이 용이함
- Shutterstock에서 라이선스된 이미지와 텍스트 데이터를 포함한 텍스트-이미지 정렬 코퍼스를 사용
- ViT-B/16 이미지 인코더를 사용

DINOv2

- DINO 작업을 따르며, 시각적 특징을 학습하기 위한 비지도 학습 방법
- 비지도 학습 방법의 주요 초점이기 때문에 선택됨
- ImageNet-22k 데이터셋을 사용하여 ViT-L 설정으로 실험을 수행

3.2 Evaluation of the Proposed Solution

Performance regression

	ImageNet Top-1	ADE20k mIoU	NYUd rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

(a) Linear evaluation with frozen features.

	ImageNet Top-1
OpenCLIP	59.9
OpenCLIP+reg	60.1

(b) Zero-shot classification.

- 레지스터 토큰을 사용함으로써 특징 표현의 품질이 영향을 받지 않는지 확인
- ImageNet 분류, ADE20k 세그먼테이션, NYU Depth V2 단일 이미지 깊이 추정에서 linear probing을 수행
- Oquab et al. (2023)의 실험 프로토콜을 따름

➡️ **레지스터를 사용했을 때 모델의 성능이 저하되지 않았으며, 때로는 더 나은 성능을 보였음**

Number of register tokens

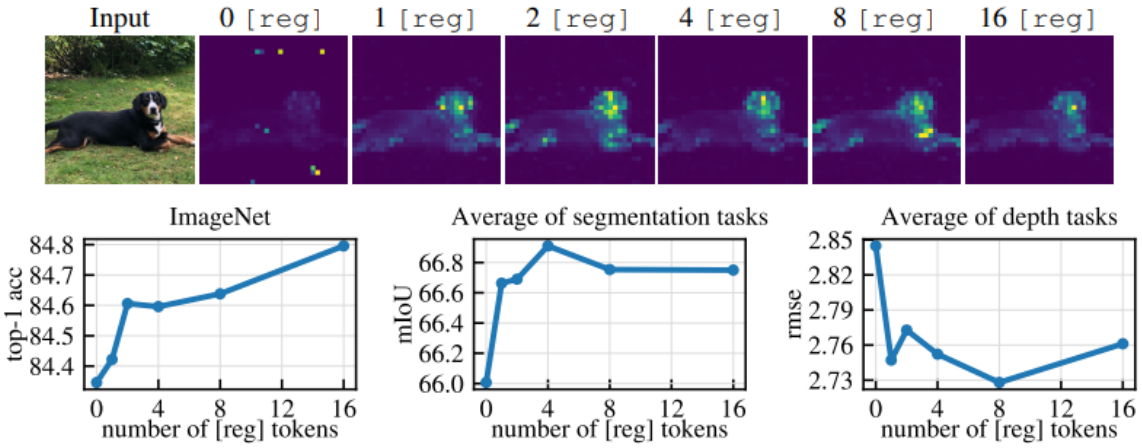


Figure 8: Ablation of the the number of register tokens used with a DINOv2 model. **(top)**: qualitative visualization of artifacts appearing as a function of number of registers. **(bottom)**: performance on three tasks (ImageNet, ADE-20k and NYUd) as a function of number of registers used. While one register is sufficient to remove artefacts, using more leads to improved downstream performance.

- 레지스터 토큰을 추가하여 특징 맵의 아티팩트를 완화하려고 함
- 이러한 토큰의 수가 로컬 특징과 다운스트림 성능에 미치는 영향을 연구
- DINOv2 ViT-L/14 모델을 0, 1, 2, 4, 8, 16개의 레지스터로 학습

➡️ **밀집 작업(dense tasks)에는 최적의 레지스터 수가 있으며, 하나를 추가하는 것이 대부분 좋은 결과를 가져옴** → 아티팩트가 사라지면서 더 나은 로컬 특징을 얻기 때문

➡️ 그러나 **ImageNet에서는 더 많은 레지스터를 사용할 때 성능이 향상됨**

3.3 Object Discovery

	VOC 2007	VOC 2012	COCO 20k
DeiT-III	11.7	13.1	10.7
DeiT-III+reg	27.1	32.7	25.1
OpenCLIP	38.8	44.3	31.0
OpenCLIP+reg	37.1	42.0	27.9
DINOv2	35.3	40.2	26.9
DINOv2+reg	55.4	60.0	42.0

Table 3: Unsupervised Object Discovery using LOST (Siméoni et al., 2021) on models with and without registers. We evaluated three types of models trained with various amounts of supervision on VOC 2007, 2012 and COCO. We measure performance using corloc. We observe that adding register tokens makes all models significantly more viable for usage in object discovery.

- 3.1의 각 알고리즘들을 사용하여 훈련되 백본에서 추출한 feature들에 대해 레지스터를 포함 혹은 미포함하여 LOST를 실행
    - PASCAL VOC 2007, 2012와 COCO 20k 데이터셋에서 객체 탐지를 수행
    - DeiT와 OpenCLIP의 경우 값을 사용하고, DINOv2의 경우 키를 사용
    - 출력 특징이 서로 다른 조건을 가질 수 있기 때문에 특징의 그램 행렬에 수동으로 바이어스를 추가함

➡️ DINOv2와 DeiT-III의 경우, 레지스터를 추가하면 **객체 탐지 성능이 크게 향상**됨

➡️ OpenCLIP의 경우, 레지스터를 사용하면 **성능이 약간 저하**됨

➡️ DINOv2의 VOC2007 성능은 DINO의 성능에는 미치지 못하지만, 레지스터를 사용한 모델은 20.1 corloc의 **성능 향상(35.3 → 55.4)**을 보였음

3.4 Qualitative Evaluation of Registers

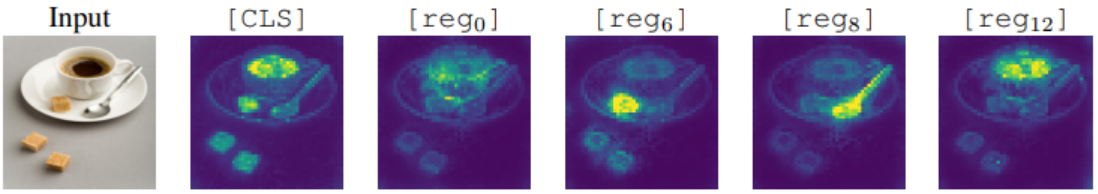


Figure 9: Comparison of the attention maps of the [CLS] and register tokens. Register tokens sometimes attend to different parts of the feature map, similarly to slot attention (Locatello et al., 2020). This behaviour was never required from the model, and emerged naturally from training.

- 레지스터 토큰들이 모두 유사한 attention 패턴을 보이는지, 아니면 자동으로 차별화가 발생하는지 확인하고자 함
- 클래스 토큰과 레지스터 토큰의 패치 토큰에 대한 attention maps을 시각화

➡️ 레지스터들을 완전히 일치된 동작을 보이지 않아쓰며, 선택된 일부 레지스터들은 scene의 다른 객체들에 주의를 기울이는 것으로 나타남

➡️ 강제한 것이 아니지만 레지스터들의 활성화는 자연스럽게 다양성을 보였으며, 레지스터의 정규화에 대한 연구는 향후 과제로 남겨짐

4. Related works

Feature extraction with pretrained models

- 2012년에 발표된 AlexNet 모델이 ImageNet-1k 데이터셋으로 사전 학습된 후, 시각적 특징을 추출하는 데 널리 사용됨
- 이후 ResNets와 같은 현대적인 아키텍처들이 등장하면서 더욱 발전되었으며 최근에는 Vision Transformers도 등장
- Transformers는 훈련 중에 다른 모달리티를 쉽게 처리할 수 있기 때문에 label supervision이나 text supervision으로 사전 학습된 백본에서 흔히 사용함
    - 물체 감지나 이미지 분할 등 다양한 작업에서 뛰어난 성능을 발휘
    - supervision은 라벨에 의존
- 자기 지도 학습
    - 감독 없이 모델이 데이터로부터 학습하도록 하는 접근법 → 사전 텍스트 작업을 설계하여 이미지의 내용을 이해하도록 함
    - Self-distillation 방법들은 고정된 백본을 사용하여 도메인 변화에 대한 더 큰 견고성을 제공함
- DINOv2
    - 본 연구에서는 자기 지도 학습, 특히 DINOv2 접근법에 초점을 맞춤
    - DINOv2는 로컬 특징 학습에 특히 효과적인 것으로 나타남
    - 그러나 뛰어난 벤치마크 점수에도 불구하고, DINOv2 특징은 바람직하지 않은 아티팩트를 나타냈음
    - 학습 과정에서 이러한 아티팩트를 수정하면 벤치마크 성능이 더욱 향상될 수 있음을 보임
- 교정 기법은 supervision 학습에서도 적용 가능(DeiT-III와 OpenCLIP 모델에서 테스트)

Additional tokens in transformers.

- Transformer 모델에서의 추가 토큰 사용
    - BERT 같은 모델에서는 SEP 토큰을 통해 네트워크에 새로운 정보를 제공하거나 CLS 토큰을 통해 분류 작업에서의 출력을 제공하는 방식으로 사용됨
    - AdaTape에서는 테이프 토큰을 추가하여 입력에 더 많은 계산 기회를 제공함
    - Memory Transformer와 같은 모델에서는 메모리 토큰을 추가하여 모델의 성능을 향상시키고, 번역 성능을 개선하는 데 활용됨

➡️ **사전 학습 단계에서 추가 토큰을 사용**하여 다양한 다운스트림 작업에서의 특징을 개선하는 것이 목표

Attention maps of vision transformers

- 비전 트랜스포머(ViT)에서의 Attention Map 시각화는 CLS 토큰과 패치 토큰 간의 어텐션 맵을 분석하는 것이 주된 방법
- DINO에서는 CLS 토큰의 attention map이 이전의 비전 트랜스포머들의 attention map과 달리 아티팩트가 없는 깨끗한 형태를 보인다고 보고하였음
- 이후 다양한 연구들이 attention map을 개선하기 위해 여러 기법들을 제안 → 비전 트랜스포머의 attention map이 더 유용하고 명확한 정보를 담을 수 있도록 함

5. Conclusion

- 본 연구에서는 DINOv2 모델의 특징 맵에서 아티팩트를 발견하고, 이러한 현상이 여러 인기 있는 기존 모델에서도 나타난다는 것을 확인했음
- 트랜스포머 모델 출력에서 아웃라이어(norm 값이 비정상적으로 높은 토큰)를 관찰하여 이러한 아티팩트를 간단히 탐지하는 방법을 제안
- 아티팩트의 위치를 연구한 결과, 모델이 정보량이 적은 영역의 토큰을 재활용하여 추론의 다른 역할로 사용하는 경향이 있음을 해석했음
- 입력 시퀀스에 추가 토큰을 첨가하는 간단한 해결책을 제안
    - 추가 토큰은 출력으로 사용되지 않으며 이를 통해 아티팩트를 완전히 제거하고 밀도 예측과 객체 발견 성능을 향상시킴
    - DeiT-III와 OpenCLIP 같은 지도 학습 모델에서도 동일한 아티팩트를 제거함을 보여줌으로써 일반적으로 적용 가능함을 확인

논문에 대한 의견 및 의문점(꼭지)

➡️ 본 논문은 레지스터 기능을 Vision Transformer 모델에 적용한 결과에 대해 설명하고 있는데, 레지스터 개념이 NLP 모델과 같이 다른 Transformer 기반 모델에도 적용될 수 있는지, 그리고 그 효과는 어떨지에 대해서도 알아보고 싶다