Vision Transformers Need Registers

지도 및 자기 지도 비전 트랜스포머 네트워크의 특징 맵에서 발생하는 artifact를 해결하기 위해 입력 시퀀스에 **추가적인 토큰**을 제공해 보자!

▼ artifact

- Attention map에서의 artifact는 모델이 이미지를 처리하는 동안 어텐션 맵에서 의도하지 않은 이상한 패턴이나 왜곡된 정보를 의미함
  - 이러한 artifacts는 모델의 해석 가능성이나 성능에 부정적인 영향을 미칠 수 있음
- attention map에서 나타날 수 있는 대표적인 artifact 유형
  1. **고노름 토큰**
     - 어텐션 맵에서 일부 토큰이 비정상적으로 높은 값을 가지는 경우
     - 이는 특정 영역이 지나치게 강조되어 다른 중요한 정보를 가리거나 왜곡할 수 있음
  2. **일관성 없는 집중(attention)**
     - 어텐션 맵이 의미론적으로 일관성 없는 영역에 집중하는 경우
     - 예를 들어, 객체의 경계를 제대로 따르지 못하고 배경 영역에 주의를 기울이는 경우
  3. **노이즈 패턴**
     - 어텐션 맵에 무작위적이거나 규칙적이지 않은 노이즈 패턴이 나타나는 경우
     - 이는 모델이 이미지의 특정 부분을 이해하지 못했음을 시사
  4. **정보 손실**
     - 어텐션 맵이 객체의 중요한 세부 사항을 놓치는 경우

        - 이는 모델이 이미지의 특정 부분에 대해 충분한 정보를 제공하지 못하는 결과를 초래할 수 있음
    - artifacts는 모델이 이미지의 의미론적 구조를 제대로 학습하지 못했거나, 모델의 내부 메커니즘에 문제가 있음을 나타낼 수 있음
        - 이는 모델의 성능을 저하시킬 뿐만 아니라, 해석 가능성 측면에서도 문제를 일으킴
        - 따라서, 이러한 artifacts를 식별하고 제거하는 것은 모델의 성능 향상과 신뢰성을 높이는 데 중요

1. Introduction

- 해당 논문은 컴퓨터 비전에서 이미지를 다목적으로 사용할 수 있는 특징으로 임베딩하는 문제를 다루고 있음
    - 초기 방법들은 수작업으로 특징을 추출했지만, 심층 학습의 발전으로 데이터 규모가 커지면서 end-to-end 학습이 가능해졌음
    - 특히, 특정 작업에 주석이 달린 데이터를 수집하는 것이 어려워서 일반적인 특징 임베딩이 여전히 중요함
        - 현재는 많은 데이터가 있는 작업에 대해 모델을 사전 학습한 후, 일부를 특징 추출기로 사용하는 방법이 일반적
- 지도 학습과 자기 지도 학습 방법 모두 높은 예측 성능과 비지도 세그멘테이션 능력으로 주목받고 있음
    - 특히 `DINO` 알고리즘은 이미지의 의미론적 정보를 잘 포착하여 객체 탐지 알고리즘인 LOST와 같은 방법에 활용될 수 있음

        ▼ LOST(**L**ocalization via **O**bject **S**egmentation with **T**ransformers)

        - 객체 탐지 알고리즘으로, 주로 `DINO`와 같은 자기 지도 학습 모델에서 얻은 특징 맵을 활용
        - 주요 특징
            - **어텐션 맵 사용**
                - DINO와 같은 모델에서 추출된 어텐션 맵을 사용하여 이미지 내의 의미론적 정보를 분석
                - 어텐션 맵은 트랜스포머 모델의 어텐션 메커니즘에서 생성된 것으로, 이미지의 특정 부분에 집중하는 경향이 있음

⇒ 이미지 내 객체의 위치와 경계를 이해하는 데 도움

- **비지도 객체 탐지**
  - LOST는 비지도 학습 방법으로, 주석이 없는 데이터를 사용하여 객체를 탐지

    ⇒ 많은 데이터를 수집하는 데 드는 비용과 노력을 줄여줌
- **의미론적 일관성 활용**
  - 어텐션 맵에서 의미론적으로 일관된 부분을 식별
  - 예를 들어, 고양이의 얼굴이나 자동차의 바퀴와 같은 특정 객체의 부분들을 인식하여 하나의 객체로 그룹화
- **밀도 기반 클러스터링**
  - 어텐션 맵에서 추출된 의미론적 정보를 기반으로 밀도 기반 클러스터링 기법을 사용하여 객체를 분할
  - 이를 통해 이미지 내에서 독립적인 객체를 분리하고 탐지할 수 있음
- **후처리**
  - 클러스터링 결과를 기반으로 추가적인 후처리 단계를 거쳐 객체의 경계를 더욱 정교하게 다듬고, 최종 객체 위치를 결정

- `DINOv2`는 `DINO`의 후속 버전으로, 밀도 있는 예측 작업에서 높은 성능을 보이지만, LOST와의 호환성에서는 기대에 미치지 못했음
  - DINOv2의 특징 맵에서 DINO에서는 없었던 artifact가 발견되었기 때문
  - 연구 결과, 이러한 아티팩트는 고노름 토큰으로, 전체 시퀀스의 작은 부분을 차지하며 중간 레이어에서 나타남

- 이를 해결하기 위해 입력 시퀀스에 **추가적인 토큰**을 도입
  - 이로 인해 artifact가 사라지고 모델의 성능이 향상되었음
  - 결과적으로 특징 맵이 부드러워져 객체 탐지 방법이 더 효과적으로 작동할 수 있게 되었음

2. Problem Formulation

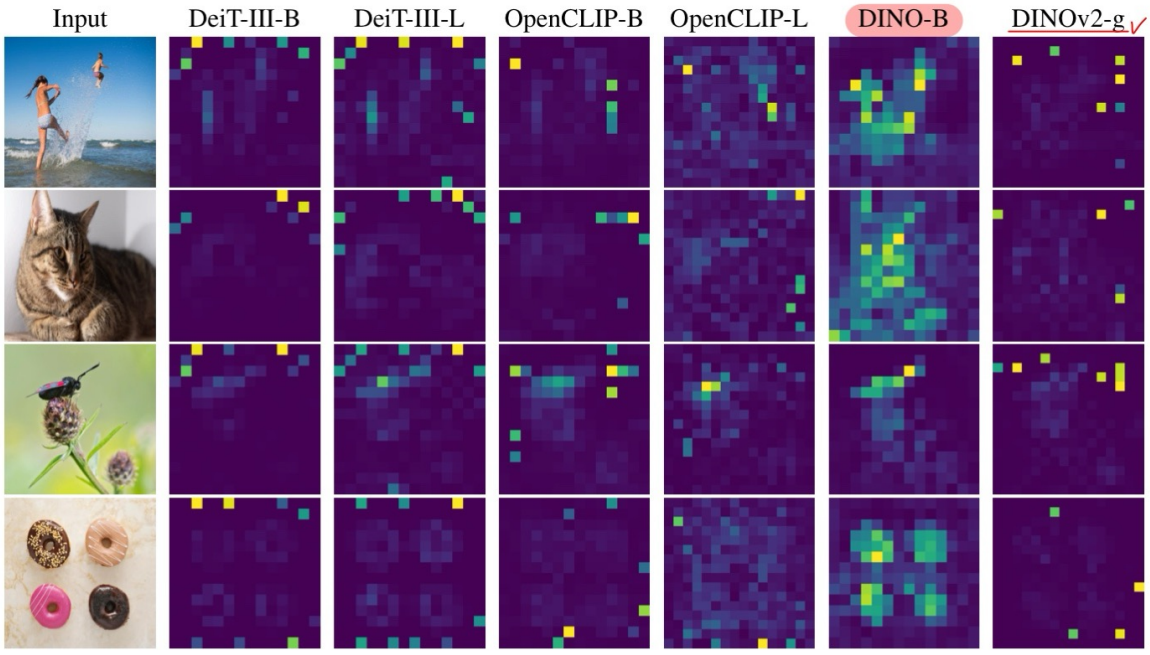


Figure 2: Illustration of artifacts observed in the attention maps of modern vision transformers. We consider ViTs trained with label supervision (DeiT-III), text-supervision (OpenCLIP) or self-supervision (DINO and DINOv2). Interestingly, all models but DINO exhibit peaky outlier values in the attention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

2-1. Artifacts In the Local Features Of DINOv2

Artifacts are high-norm outlier tokens

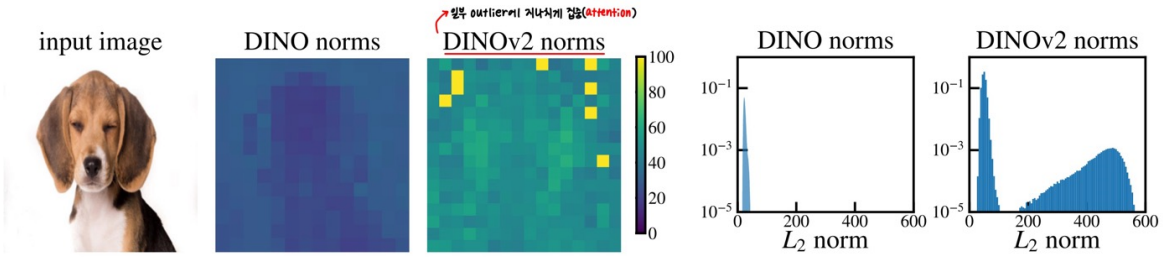
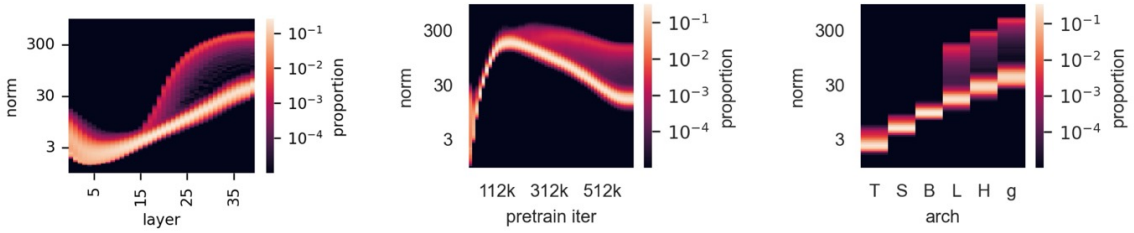


Figure 3: Comparison of local feature norms for DINO ViT-B/16 and DINOv2 ViT-g/14. We observe that DINOv2 has a few outlier patches, whereas DINO does not present these artifacts. For DINOv2, although most patch tokens have a norm between 0 and 100, a small proportion of tokens have a very high norm.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

- 아티팩트 패치와 다른 패치의 중요한 차이점은 모델 출력 시 토큰 임베딩의 **노름(norm)** 임
  - `DINO`와 `DINOv2` 모델의 로컬 특징 노름을 비교한 결과, 아티팩트 패치의 노름이 다른 패치보다 훨씬 높음

- 작은 데이터셋의 특징 노름 분포는 이중 봉우리(bimodal)로 나타나므로, 노름이 150을 초과하는 토큰을 "고노름" 토큰으로 간주하고 이들의 특성을 연구

Outliers appear during the training of large models

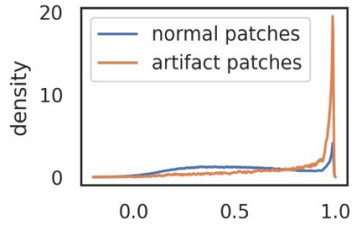


(a) Norms along layers. (b) Norms along iterations. (c) Norms across model size.

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.

- 큰 모델의 학습 중에 이상치 패치가 나타남
- DINOv2 학습 중 고노름 패치는 15층 주위에서 다른 패치와 구분되며, 학습의 3분의 1이 지난 후에 나타남
    - 또한, 크기가 큰 모델에서만 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.

- 고노름 토큰은 정보가 **중복**되는 곳에서 나타남
- 고노름 토큰과 이웃하는 토큰 사이의 코사인 유사도를 측정한 결과, 고노름 토큰이 이웃하는 패치와 매우 유사한 패치에서 나타나는 것을 확인했음

- 모델이 이미지 표현의 품질을 해치지 않고도 이러한 정보를 버릴 수 있음을 시사

High-norm tokens hold “little” local information

- 고노름 토큰은 지역 정보를 거의 포함하지 않음
  - 위치 예측과 픽셀 재구성 작업에서 고노름 토큰이 다른 토큰보다 낮은 정확도를 보였음
  - 이는 고노름 토큰이 위치 정보나 이미지 재구성에 필요한 정보를 덜 포함하고 있음을 시사

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.

- 고노름 토큰은 전역 정보를 포함
  - 이미지 표현 학습 벤치마크에서 고노름 토큰을 평가한 결과, 고노름 토큰이 다른 토큰보다 높은 정확도를 보였음
  - 이는 이상치 토큰이 다른 패치 토큰보다 더 많은 전역 정보를 포함하고 있음을 시사

2-2. Hypothesis And Remediation

- 해당 연구를 통해 대규모 비전 트랜스포머 모델들이 훈련 과정에서 중복된 정보를 인식하고 이를 전역 정보를 저장하는 용도로 사용하는 경향이 있다는 가설을 제안
  - 이로 인해 모델은 로컬 이미지 패치 정보를 희생할 수 있고, 이는 밀도 예측 작업 등에서 성능 저하를 초래할 수 있음
- 이를 해결하기 위해, 토큰 시퀀스에 **추가적인 레지스터 역할**을 할 수 있는 **새로운 토큰을 도입**하는 간단한 방법을 제안
  - 이는 비전 트랜스포머의 해석 가능성과 성능 문제를 개선하는 것으로 나타났음
  - 또한, 다양한 모델에서 나타나는 아티팩트의 원인은 학습 방법과 모델 크기, 훈련 기간 등 여러 요소가 복합적으로 작용하는 것으로 추론됨

3. Experiments

3-1. Training Algorithms and Data

- **지도 학습, 텍스트 지도 학습, 비지도 학습**의 세 가지 최신 훈련 방법에 적용

1. **DEIT-III (Touvron et al., 2022)**
    - DEIT-III는 ImageNet-1k 및 ImageNet-22k에서 ViTs를 사용한 간단하고 견고한 지도 학습 분류 방법임
    - 이 방법은 베이스 ViT 구조를 사용하며 강력한 분류 결과를 달성하며, 공식 저장소에서 제공하는 ViT-B 설정을 사용하여 ImageNet-22k 데이터셋에 적용
2. **OpenCLIP (Ilharco et al., 2021)**
    - OpenCLIP은 원본 CLIP 작업을 따라 텍스트-이미지 정렬 모델을 생성하는 강력한 텍스트 지도 학습 방법임
    - 이 방법은 오픈 소스이며 베이스 ViT 구조를 사용하며, Shutterstock에서 라이센스된 이미지 및 텍스트 데이터를 기반으로 한 텍스트-이미지 정렬 말뭉치에서 실행됨
    - 공식 저장소에서 제안하는 ViT-B/16 이미지 인코더를 사용
3. **DINOv2 (Oquab et al., 2023)**
    - DINOv2는 DINO 작업을 따르는 자기 지도 학습 방법으로, 시각적 특징을 학습
    - 이 방법은 연구의 주요 관심 대상으로, 공식 저장소에서 제공하는 ViT-L 구성을 사용하여 ImageNet-22k에서 실행됨

3-2. Evaluation Of The Proposed Solution

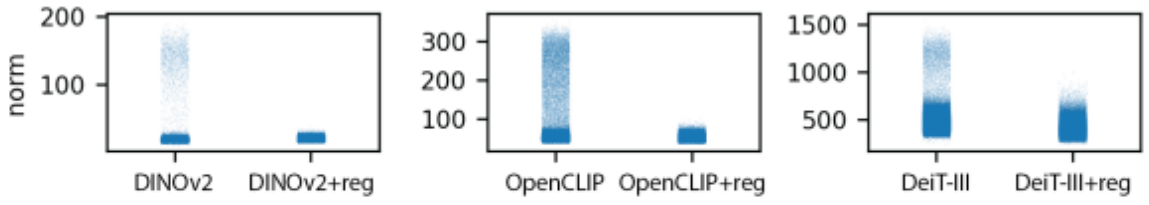


Figure 7: Effect of register tokens on the distribution of output norms on DINOv2, OpenCLIP and DeiT-III. Using register tokens effectively removes the norm outliers that were present previously.

- 훈련할 때 레지스터(= 추가적인 token)를 사용하면 모델이 출력에서 큰 norm 토큰을 보이지 않음을 확인할 수 있음

Performance regression

- 레지스터 토큰을 사용하면 local feature map에서 artifact를 제거함을 확인
  - 이러한 특성이 품질에 영향을 미치지 않는지 확인
- 레지스터 토큰을 사용할 때 모델은 성능을 잃지 않고 때로는 더 나은 결과를 보임을 확인

	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.

Table 2: Evaluation of downstream performance of the models that we trained, with and without registers. We consider linear probing of frozen features for all three models, and zero-shot evaluation for the OpenCLIP model. We see that using register not only does not degrade performance, but even improves it by a slight margin in some cases.

Number of register tokens

- 특성 맵의 아티팩트를 완화하기 위해 레지스터 토큰을 추가하는 것을 제안
  - 해당 실험에서는 이러한 토큰 수가 로컬 특성 및 하향 평가 성능에 미치는 영향을 연구

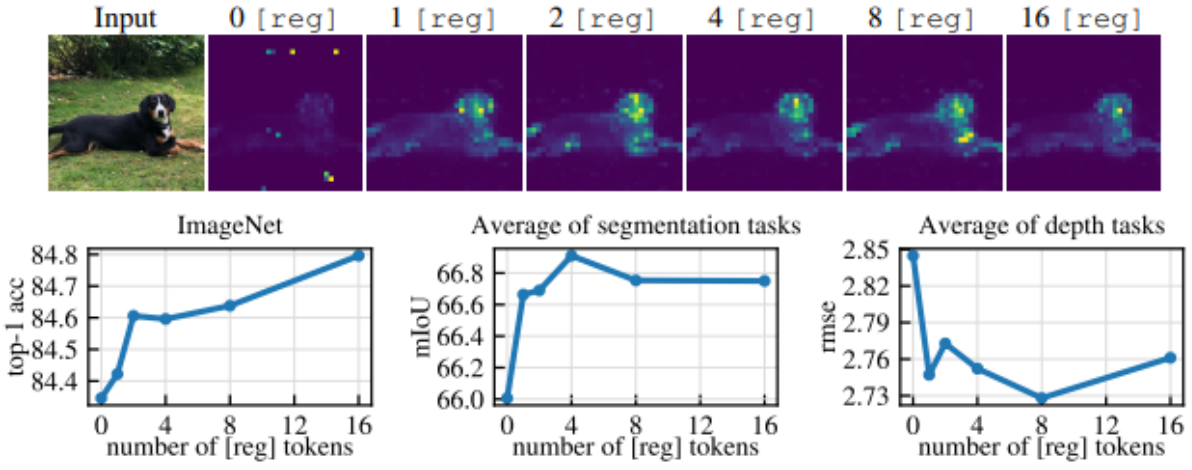


Figure 8: Ablation of the the number of register tokens used with a DINOv2 model. **(top)**: qualitative visualization of artifacts appearing as a function of number of registers. **(bottom)**: performance on three tasks (ImageNet, ADE-20k and NYUd) as a function of number of registers used. While one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

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.

- 최근 비지도 객체 발견 방법들은 DINO를 기반으로 하여 로컬 특성 맵의 품질과 부드러움에 크게 의존함
  - 그러나 최신 백본인 DINOv2나 DeiT-III와 같은 모델에 이 방법을 적용할 때 성능이 저하되는 문제가 발생
- 본 연구에서 제안하는 방법은 이 문제를 완화할 수 있다고 주장하며, 이를 검증하기 위해 다양한 데이터셋에서 실험을 수행하였음
  - 실험 결과, DINOv2와 DeiT-III에서는 레지스터를 추가하여 객체 발견 성능이 크게 향상되었으며, OpenCLIP에서는 성능이 약간 저하되는 경향을 보였음
  - 특히, DINOv2는 VOC2007에서 이전 연구에서 보고된 DINO의 성능에는 미치지 못하지만, 레지스터를 추가한 모델은 bet corloc 55.4로 20.1 corloc의 개선을 보였음

3-4. Qualitative Evaluation Of Registers

- 모든 레지스터 토큰이 유사한 attention 패턴을 보이는지 아니면 자동으로 차이가 나타나는지 확인
  - 이를 위해 클래스 및 레지스터 토큰의 주의 맵을 패치 토큰에 대해 시각화하여 플롯

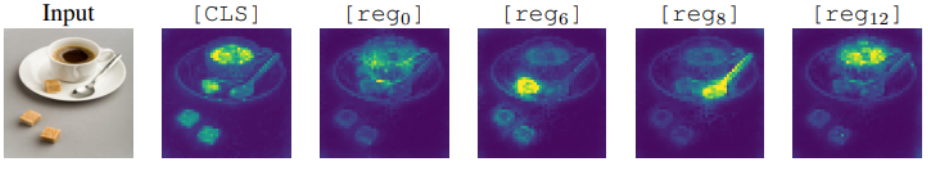


Figure 9: Comparison of the attention maps of the [CLS] and register tokens. Register tokens sometimes attend to different parts of the feature map, similarly to slot attention (Locatello et al., 2020). This behaviour was never required from the model, and emerged naturally from training.

- 레지스터가 완전히 일치하는 동작을 보이지 않는다는 것을 확인할 수 있음
    - 일부 선택된 레지스터는 재미있는 주의 패턴을 보이며, 장면의 다른 객체에 주의를 기울임

4. Related Work

Feature extraction with pretrained models

- 사전 학습된 신경망 모델은 AlexNet 이후로 시간이 경과함에 따라 이미지넷-1k 데이터셋에서 학습한 CNN 모델을, 최근에는 ResNet 등 현대 아키텍처가 사용됨
- 비전 트랜스포머는 다양한 모달리티를 처리할 수 있어, 레이블 지도나 텍스트 지도로 훈련된 백본을 통해 강력한 시각적 기초 모델을 제공하며 다양한 작업에서 우수한 성능을 보임

Additional tokens in transformers

- 비지도 학습은 데이터 이해를 위한 전제 작업을 통해 모델이 학습됨
- 비전 트랜스포머를 사용한 자기 지도 학습 방법은 MAE와 자기 증류 계열 방법을 포함하며, 이는 고정된 백본을 사용하여 강력한 성능을 보이며 도메인 변화에 강건함

Attention maps of vision transformers

- DINO를 통해 주의 맵 시각화가 인기를 얻었으며, DINO는 깨끗한 주의 맵을 보여줌
- 최근 연구에서는 다양한 기법을 사용하여 흥미로운 주의 맵을 보고, 최적화 절차 수정, 유용한 이미지 부분으로 주의 점수 조정, 트랜스포머 레이어 아키텍처 수정, 학습 가능한 풀링 도입 등이 포함됨

5. Conclusion

- 해당 연구에서는 DINOv2 모델의 특성 맵에서 발생하는 artifact를 밝혔으며, 이 현상이 여러 인기 있는 모델에서도 관찰된다는 것을 발견했음
    - Transformer 모델의 출력에서 이상치(norm 값이 매우 높은 토큰)에 해당하는 것들을 관찰함으로써 이러한 artifact를 감지하는 간단한 방법을 설명
    - 이들의 위치를 연구하여, 모델이 정보가 적은 영역에서 토큰을 재활용하고 이를 추론 과정에서 다른 역할로 재사용하는 것으로 해석
        - 이 해석을 바탕으로, 입력 시퀀스에 출력으로 사용되지 않는 추가 토큰을 추가하는 간단한 수정 방법을 제안했고, 이로써 artifact를 완전히 제거하고 밀집 예

측과 객체 발견 작업에서 성능을 향상시킬 수 있음을 발견했음

- 더불어, 제안된 해결책이 DeiT-III나 OpenCLIP 같은 지도 학습 모델에서도 동일한 artifact를 제거함을 보여줌으로써 이 해결 방법의 일반성을 확인했음