[11주차] 논문리뷰

MAE

0. Abstract

- MAE가 CV 분야에서 scalable self-supervised learners임을 증명했음
- 아이디어 : NLP에서 데이터의 일부를 지우고, 지운 내용들을 예측하도록 모델을 학습하는 것과 비슷한 방식을 CV에 적용
- 입력 이미지의 패치를 랜덤하게 마스킹한 후 missing pixels을 복원하도록 학습

1. Introduction

논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가

논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

논문이 다루는 분야

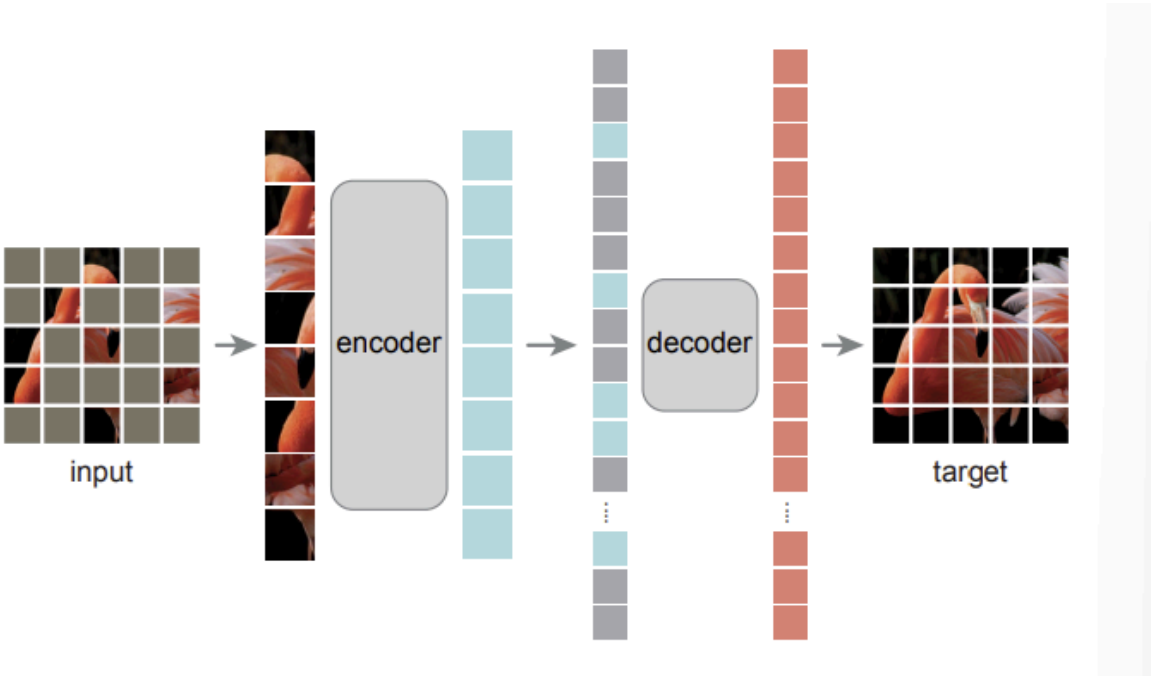
- Computer Vision (autoencoding methods in vision)
- visual representation learning
- NLP에서 성공한 방식을 CV에 적용하려고 했음
  - NLP 성공 모델 : GPT(autoregressive 언어 모델), BERT(masked autoencoding 모델- denoising autoencoders의 일반화된 형태)
  - 이들의 아이디어 : 데이터의 일부를 지우고, 지운 내용을 예측하도록 모델 학습
- *"무엇이 NLP와 CV에서 autoencoding method의 적용을 다르게 만드는 지점일까?"*

  1. Architecture가 다르다
     - CV에서는 보통 CNN 사용하는데, 이는 NLP와 달리 indicator(mask token, positional embedding 등)가 존재하지 않는다 → ViT의 등장으로 해결!
  2. Information density가 다르다
     - Languages는 의미가 풍부하고 정보 밀도가 높다
       → 단어 하나만 빠져도 문맥 이해 복원이 어렵다
       → 단어 예측하려면 자연스럽게 정교한 언어 이해 능력을 얻게 됨
     - Images는 공간 정보가 중복되어있음
       → 일부를 가려도 주변만 봐도 해당 부분을 잘 예측할 수 있음
       → 고차원적 의미 이해가 필요가 없음
     - 해결법 : 더 많이 가린다! (=더 높은 비율로 랜덤하게 패치를 마스킹한다)
       - 일단 공간적 중복성을 감소시켜준다.

        - 텍스트의 경우 15%만 가려도 모델이 의미를 이해하는데, 이미지의 경우 75% 이상을 가려야 단순 복원이 아닌 전체 구조, 장면에 대한 의미를 이해하기 시작한다는 의미.
        - = 언어처럼 의미있는 복원이 이루어지려면 더 많이 가려야 한다는 것.
    3. Autoencoder의 Decoder가 수행하는 바가 다르다
        - CV: Pixels를 재구축함. 정보의 의미수준이 낮음.
            - 일반적인 비전 task는 semantic level이 높기 때문에 픽셀 복원만 잘하는 건 의미가 없음
        - NLP: missing *words*를 예측함. 단어는 풍부한 semantic 정보를 포함함. 자연스럽게 semantic level이 높은 학습이 이루어짐.
        - BERT(NLP)에서는 디코더가 단순한 MLP여도 충분하지만, 이미지에서는 어떤 디코더 구조냐에 따라 semantic level이 크게 좌우됨. 디코더를 정교하게 설계해야 의미있는 latent representation을 만들어낼 수 있음.
- 위와 같은 분석을 바탕으로 간단하고, 효과적이고, 확장 가능한 형태의 모델을 제안함.

해당 task에서 기존 연구 한계점

논문의 contributions

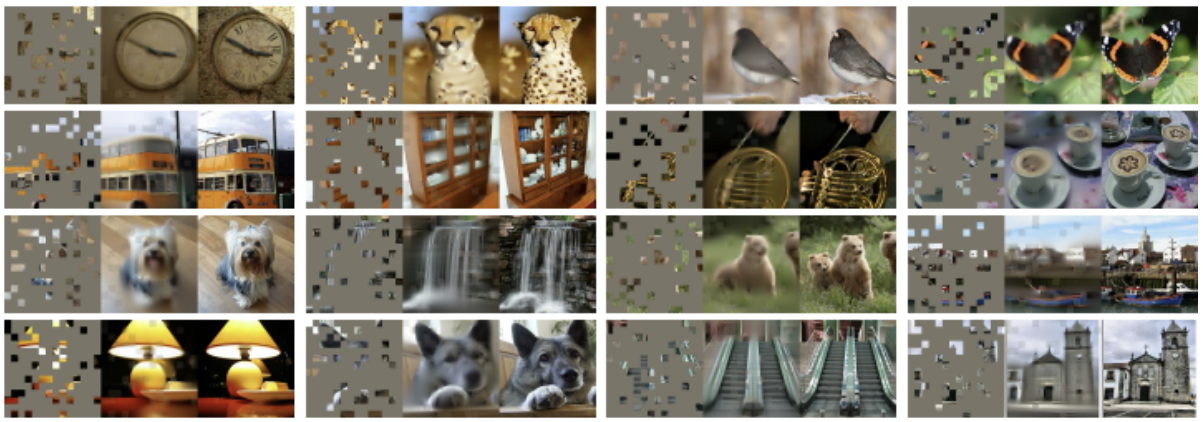


- 인코더-디코더 구조는 비대칭 구조

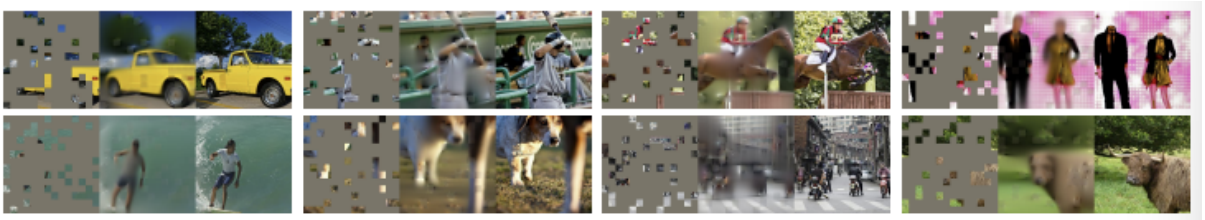
    - 인코더: 마스킹 되지 않은 부분만 인코딩함
    - 디코더:
        - 인코더보다 훨씬 가볍게 설정
        - 마스킹 된 부분과 되지 않은 부분 모두 디코딩함
    - 최적 마스킹 비율은 75%임을 찾아냄
    - 메모리 연산량을 줄였음. 전체적인 학습속도도 줄임. → 큰 모델로 확장할 수 있음
- MAE는 high-capacity 모델도 잘 일반화시킬 수 있는 사전학습 기법
    - ViT-Large, ViT-Huge 같이 데이터를 많이 요구하는 모델
      → ImageNet-1K에서 더 나은 일반화 성능으로 학습 가능
    - ViT-Huge 모델을 사용해 ImageNet-1K에서 파인튜닝한 결과, 87.8%의 정확도를 달성
- 전이 학습 성능도 검증했음
    - object detection, instance segmentation, semantic segmentation
    - MAE 기반 사전 학습이 기존의 지도 학습 기반 사전 학습보다 더 우수한 성능
    - 모델의 크기를 키웠을 때 성능이 크게 향상
- NLP에서 BERT류 모델들이 보여준 발전과 비슷하며, 비전 분야에서도 비슷한 발전을 기대할 수 있음!!

[가시적인 결과]

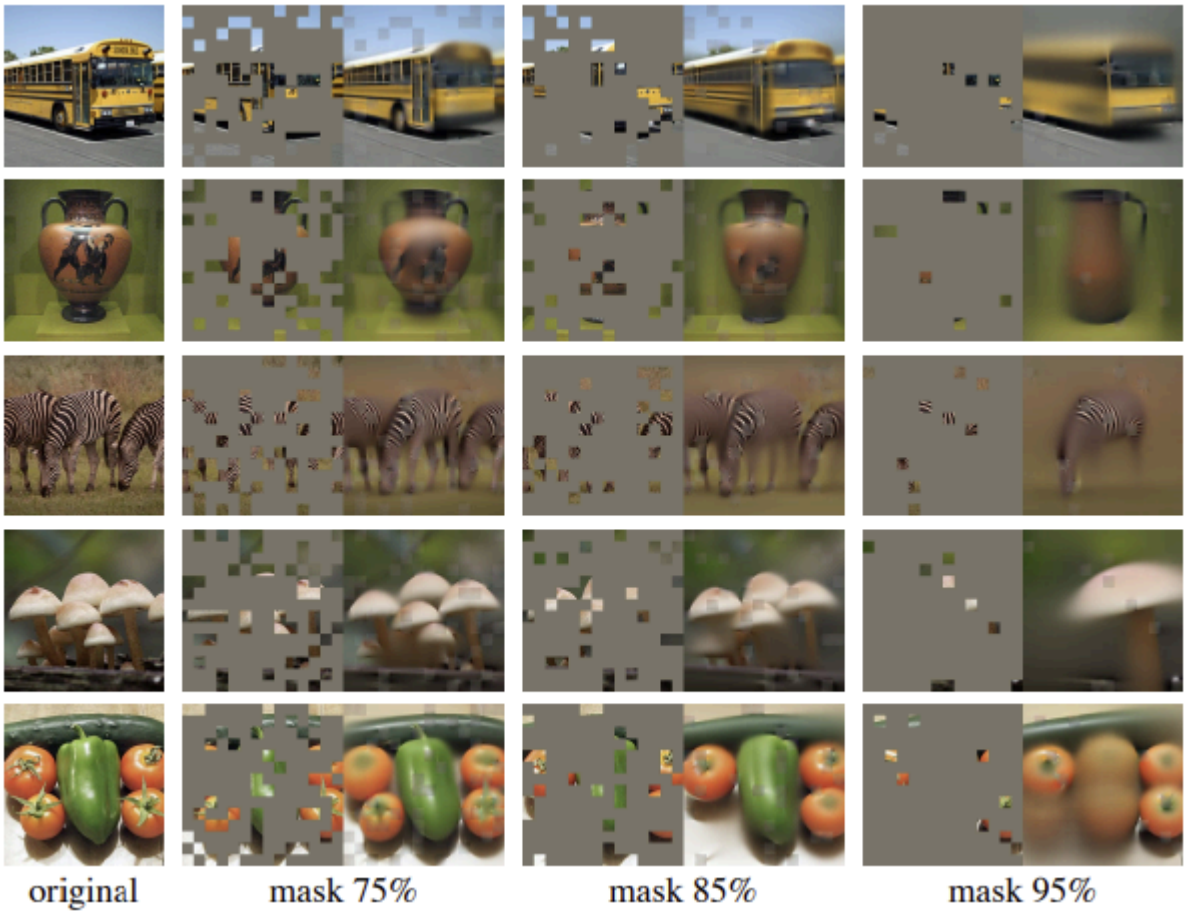
- ImageNet 데이터셋으로 학습된 MAE 사용
- 입력의 대부분(80%)이 가려졌음에도 복원결과(중간)가 원본(오른쪽)과 매우 비슷함.
- 단순한 복원이 아닌 의미 있는 high semantic level의 복원이 이루어졌음을 확인할 수 있음.



- Transfer learning의 성능을 보여주고자 COCO dataset의 validation 이미지에서 test함
- 모델은 위에서 ImageNet 데이터셋으로 학습된 MAE 똑같이 사용
- 여기서도 단순한 복원이 아닌 의미 있는 high semantic level의 복원이 이루어졌음을 확인할 수 있음.



- 마스킹 비율 변화에 따른 복원 결과의 성능
- 학습 자체는 75%에서 하고, test를 75%, 85%, 95%에서 진행해서 마스킹 비율에 따라 성능 비교.
- 학습 때보다 더 많이 가려도 나름 그럴듯하게 복원함.



2. Related Work

💡 Introduction에서 언급한 기존 연구들에 대해 어떻게 서술하는가
제안 방법의 차별성을 어떻게 표현하고 있는가

1. **Masked language modeling(MLM)**
   - NLP에서의 성공적인 사전학습 모델인 BERT와 GPT가 대표적
   - 입력 시퀀스의 일부를 지우고, 지워진 부분을 재복원할 수 있도록 학습하는 전략

   ⇒ MLM은 언어의 정보 밀도가 높아 일부만 가려도 충분한 의미 학습이 가능한 반면 이미지는 공간적으로 중복이 많아, 높은 마스킹률을 적용해야 의미 기반 복원이 유도됨.
2. **Autoencoding**
   - 인코더에서 입력 데이터를 잠재 표현에 매핑하고, 디코더에서 재복원하는 모델
   - DAE : 입력 신호를 망가뜨리고, 재복원 시에 온전한 신호로 복원하는 방법

    - 픽셀 마스킹하거나 색상 제거하는 것도 일종의 DAE

⇒ MAE는 기존 autoencoder와 달리 asymmetric design을 채택해서 인코더는 보이는 patch만 인코딩하고, 디코더는 전체 patch로 복원하게 함. 이는 고효율 훈련을 가능하게 하고, ViT-Huge 같은 대규모 모델 학습을 가능하게 함.

3. **Masked image encoding**
    - 마스킹에 의해 망가뜨려진 이미지로부터 표현을 추출하는 방법(마스킹=노이즈)
    - 초기연구 : iGPT, BEiT

⇒ 기존 방식들은 보통 encoder에 전체 이미지를 입력하는 반면, MAE는 encoder는 일부 patch만 처리하여 훈련 효율성과 확장성을 높임.

⇒ BEiT 등은 masked token 예측을 위해 discrete visual tokens (VQ-VAE 등) 필요한 반면, MAE는 pixel-level 복원을 사용하므로 추가 tokenizer 없어도 됨.

4. **Self-supervised learning**
    - 사전학습을 위해 여러 pretext task 집중하는 방법
    - 대조학습 : 오토인코더를 활용하는 것과는 개념적으로 다른 방향을 추구함. 증강에 민감

⇒ MAE는 대조학습과 달리 explicit positive/negative 쌍 없이도 복원 기반의 자기지도학습 가능

3. 제안 방법론

Introduction에서 언급된 내용과 동일하게 작성되어 있는가

Introduction에서 언급한 제안 방법이 가지는 장점에 대한 근거가 있는가

제안 방법에 대한 설명이 구현 가능하도록 작성되어 있는가

Main Idea

💡 부분적으로 관측된 입력을 기반으로 원래의 신호를 복원하는 간단한 오토인코더.

공통점

- 인코더 : 입력을 잠재 표현으로 매핑
- 디터더 : 잠재 표현으로부터 원래 신호를 복원

차이점

- 비대칭적 인코더-디코더 구조를 가짐
- 인코더가 마스킹되지 않은 패치들만 처리함
- 디코더는 인코더에 비해 가벼운 구조임

1. **Masking**
   - ViT 방식에 따라, 겹치지 않게 패치들을 나누고, 랜덤하게 마스킹한 후 마스킹되지 않은 것들을 입력으로 사용함. 높은 마스킹 비율을 적용. 마스킹된 패치는 제거됨
   - 랜덤 샘플링을 통해 중앙 패치들이 과도하게 마스킹되는 현상을 방지
2. **MAE encoder**
   - ViT와 동일한 인코더 구조이지만, visible한 패치들(25% 정도)에만 적용됨.
   - 각 입력 패치는 선형 투사 후 positional embedding이 더해진 뒤, 트랜스포머 블록을 거쳐 처리됨
   - 계산량이 크게 줄고, 대형 모델도 효율적으로 학습할 수 있게 됨.
3. **MAE decoder**
   - 사전학습(pre-training) 단계에서만 사용. 이미지 복원을 위한 목적.
     - 실제로 모델을 이미지 분류, 탐지, 분할 같은 downstream task 등에 사용할 때는 모델의 decoder는 필요하지 않음. decoder의 역할은 이미지를 복원하는 훈련을 통해 encoder가 더 좋은 표현을 배우게 하는 것. 그 이후에는 encoder만 써서 이미지 분류나 전이학습 등에 사용함.
   - 다양한 경량 디코더를 실험했음. 인코더보다 작고 얕고 토큰당 계산량이 인코더의 10% 이하
4. **Reconstruction target**
   - MAE는 입력 이미지를 픽셀 수준에서 복원

- 디코더 출력은 각 패치들의 픽셀 값들을 나타내는 벡터. 마지막 디코더 층은 픽셀 개수와 동일한 채널 수를 가짐
- 출력은 재구성된 이미지로 reshape됨
- 손실함수는 원본이미지와 복원이미지 간 MSE로 정의됨. 손실은 마스킹 된 부분에 대해서만 계산
- “normalized pixel”를 복원 대상으로 삼는 것도 실험함
    - 각 패치의 평균과 표준편차를 이용해 정규화한 값만 복원하게 함 → 학습표현의 품질이 향상됨

5. **Simple implementation**
    - 복잡한 sparse 연산 없이도 효율적으로 구현 가능
    1. 선형 투사 + positional embedding을 통해 각 패치를 토큰으로 변환
    2. 리스트를 무작위로 섞어 일정 비율을 제거해서 보이는 토큰 리스트를 생성
    3. 이후 인코딩된 토큰 리스트에 mask token을 다시 추가
    4. 리스트를 원래 순서에 맞게 재정렬(shuffle back)해서 디코더 입력으로 사용

Contribution

- 전반적으로 계산량이 크게 줄어서 더 큰 모델도 학습할 수 있게 됨

4. 실험 및 결과 (ImageNet Experiments)

Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

Dataset

- ImageNet-1K / Self-supervised learning
- (i) 전체 fine-tuning 또는 (ii) linear probing을 사용해 감독학습(fine-tuning)을 수행
- 224×224 크롭 단일 이미지를 기준 / top-1 validation accuracy

Baseline

- backbone : ViT-Large (ViT-L/16) (ablation study)

- baseline MAE

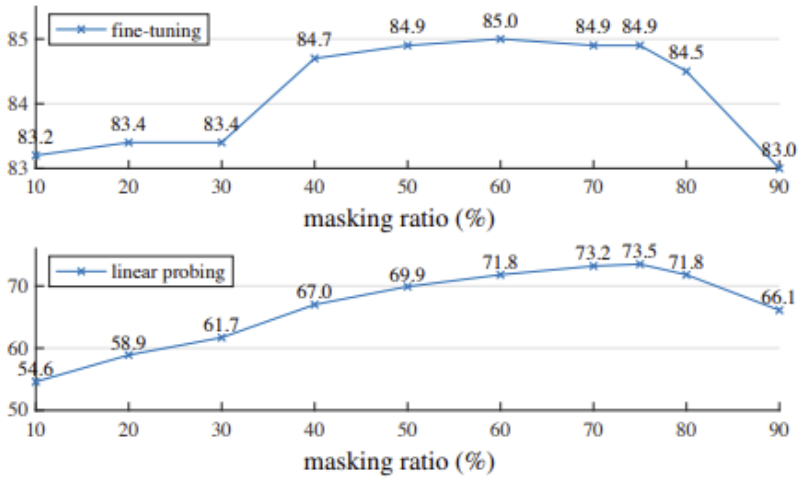
결과

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

- ViT-L을 scratch로 학습하는 것은 쉽지 않음. MAE pre-training은 이보다 훨씬 좋은 성능을 냄.
- MAE 사전학습이 fine-tuning 성능에 매우 큰 영향을 준다는 것을 의미

Main Properties

- **마스킹 비율**



  - 75%의 마스킹 비율이 fine-tuning과 linear probing 모두에서 가장 좋은 결과
    - fine-tuning은 마스킹 비율에 덜 민감. 40~80% 범위에서 모두 높은 성능
    - linear probing은 성능이 마스킹 비율에 따라 점진적으로 증가. 20% vs 75%에서 약 20%p의 성능 차이
    - 모든 결과는 scratch 학습보다 훨씬 우수

- **디코더 설계**

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

- 유연하게 설계, encoder와 분리되어 설계
- linear probing에는 디코더의 깊이가 중요
  - 깊은 디코더는 reconstruction에는 유리하나, recognition에는 덜 적합
- fine-tuning에서는 encoder가 recognition task에 적응하므로, 디코더 깊이의 영향이 적음
- 단 하나의 transformer block으로 구성된 디코더도 fine-tuning에서 84.8% 성능을 기록

- **마스크 토큰 설계**

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

- MAE 핵심 설계 : mask token을 입력하지 않는 것
- mask token 넣으면 linear probing 성능이 14% 가량 저하됨
- mask token을 제거하면, 인코더는 real 패치만 보고 학습하게 되어 일반화 성능이 향상
- mask token 제거로 인해 연산량(FLOPs)이 3.3× 감소, 학습 속도는 최대 4×까지 빨라짐

- **복원 목표 (Reconstruction target)**

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

- 기본적으로는 정규화되지 않은 픽셀 복원 대상으로 사용
- patch-wise하게 정규화된 픽셀을 사용하면 성능이 더 좋아짐(PCA 기반 복원 실험은 성능 더 나빠짐)
  - patch 내부의 평균과 표준편차로 정규화
- BEiT처럼 token을 복원 대상으로 설정한 실험도 수행
  - fine-tuning 성능은 0.4% 향상되었으나, linear probing 성능은 오히려 감소
  - 픽셀 복원이 더 단순하고 효과적임을 의미함

- **데이터 증강**

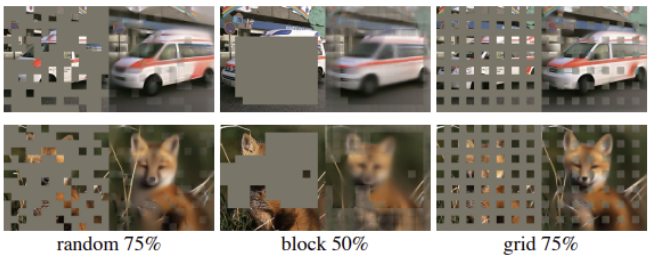
case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

- contrastive learning처럼 강력한 data augmentation 없이도 좋은 성능을 냄
- cropping-only 전략(중앙 or 랜덤 크롭)만으로도 충분함
  - cropping-only : 다른 복잡한 증강 없이 이미지를 자르는 방법 - 중앙 자르기나 랜덤 자르기
- flipping, color jitter 등을 추가하면 성능이 오히려 감소
  - flipping : 이미지를 좌우 또는 상하로 뒤집는 것. 일반적으로 의미가 바뀌지 않는 객체에 유용
  - color jitter : 색상, 밝기, 대비, 채도 등을 무작위로 변화시키는 증강 기법. 색상 정보에 의존X도록

- contrastive 방식처럼 다양한 뷰(view)를 만들 필요가 없기 때문에 간단한 증강만으로도 효과적

- **마스크 샘플링 전략**



case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

  - random sampling > block-wise > grid-wise
  - random masking이 복원 task를 어렵게 만들어서 encoder가 더 나은 표현을 학습하게 함.
  - block-wise는 50%에서 괜찮지만 75%에서 성능 저하
  - grid-wise는 쉬운 복원이 가능해서 표현 학습에는 부적합

- **학습 스케줄**

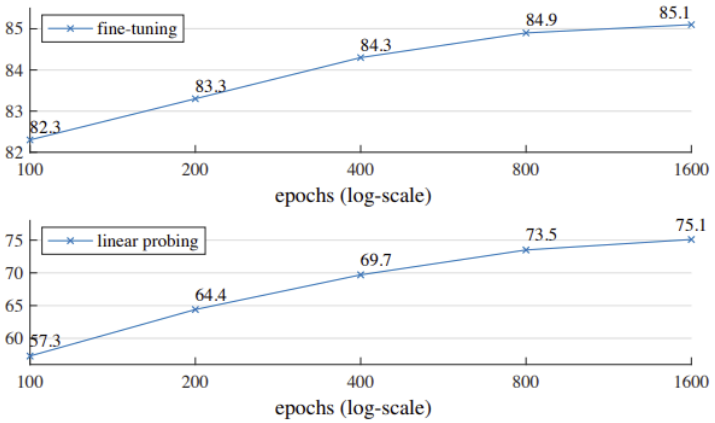


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

Table 2. **Wall-clock time** of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

- 800 epoch 사전학습을 기본, 더 길게 학습하면 성능이 계속 증가
- linear probing은 1600 epoch까지도 성능 향상
- MAE는 다른 방법보다 학습 시간도 더 짧음
  - ViT-L을 128 TPU v3-core에서 학습할 때: MAE는 31시간, MoCo v3는 36시간

기존 방법과의 비교

1. **Comparisons with self-supervised methods**

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

- self-supervised methods 기반 ViT 모델들의 fine-tuning 결과를 비교
- MAE는 모델 크기를 쉽게 확장할 수 있음과, 더 큰 모델에서의 지속적인 성능 향상을 보여주었음
- BEiT(토큰 복원, 사전 훈련된 dVAE 토크나이저 필요)와 비교하면, 우리의 MAE(픽셀 복원)는 더 정확하면서도, 구조가 더 단순하고 훨씬 빠름, epoch당 학습 속도도 BEiT보다 3.5배 빠름

2. **Comparisons with supervised pre-training**

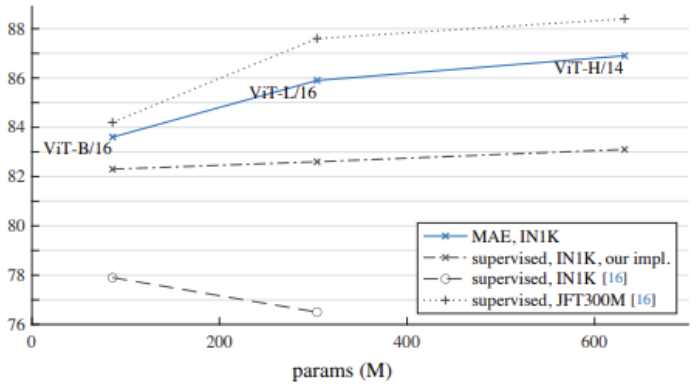


Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

- 원래 ViT 논문에서는, ViT-L이 IN1K에서 학습될 경우 성능이 저하되는 것으로 나타났음
- MAE의 사전 학습은 IN1K 데이터만을 사용했음에도 불구하고 더 나은 일반화 능력을 보여주었음
- 모델 크기가 클수록, scratch로 학습했을 때보다 MAE를 활용한 학습이 더 큰 성능 향상을 가져옴
- 이 비교는 MAE가 대형 모델로 확장하는 데 유리하다는 점을 보여줌

Partial Fine-tuning

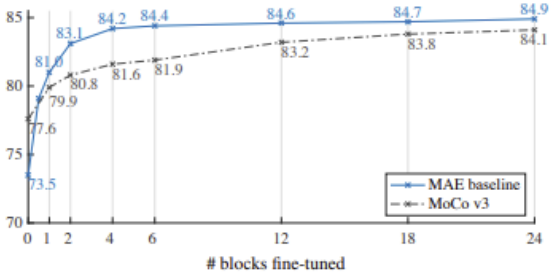


Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

- Partial Fine-tuning : 마지막 몇 개의 레이어만 fine-tune하고 나머지는 고정
- 단 하나의 Transformer 블록만 fine-tuning 해도, 정확도가 73.5%에서 81.0%로 크게 향상함
- 마지막 블록의 일부(예: MLP sub-block)만 fine-tune해도 79.1%의 정확도

⇒ 몇 개의 블록만 fine-tune해도 전체 fine-tuning 수준의 성능에 도달할 수 있음

- MoCo v3는 ViT-L 결과가 공개된 contrastive 방법
  - 더 높은 전체 fine-tuning 정확도를 갖지만
  - Partial Fine-tuning 실험에서는 모두 MAE보다 성능이 낮았음

⇒ MAE 표현은 비선형적 표현에서 더 강력하며, 비선형 분류기를 사용할 때 매우 잘 작동함

⇒ linear separability 만으로는 표현 품질을 평가하기에 충분하지 않다

5. 실험 및 결과 (Transfer Learning Experiments)

💡 Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

Model

- Table 3의 사전학습된 모델들을 사용
- 다운스트림 작업에서의 전이 학습 성능을 평가

Object detection and segmentation

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	44.9	47.2

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

- COCO 데이터셋
- Mask R-CNN을 end-to-end로 fine-tune
- ViT 백본은 FPN구조에 맞춰 조정
- 객체 탐지(APbox) 및 인스턴스 분할(APmask)을 기준으로 평가

- supervised pre-training과 비교했을 때, MAE 사전학습 모델이 모든 구성에서 더 나은 성능
- pixel 기반 MAE는 token 기반 BEiT보다 우수하거나 동등한 성능을 보임

Semantic segmentation

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

- ADE20K 데이터셋
- UPerNet을 사용

- MAE 사전학습이 감독 학습보다 훨씬 높은 성능을 달성함
- pixel 기반 MAE는 token 기반 BEiT보다 더 나은 성능을 보였음.

Classification tasks

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [55]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [54]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [54]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [40]‡

Table 6. **Transfer learning accuracy on classification datasets**, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.
†: pre-trained on 1 billion images. ‡: pre-trained on 3.5 billion images.

- iNaturalist와 Places 데이터셋
- 전이 학습 결과를 비교

- MAE는 모델이 클수록 성능이 향상되는 scaling behavior(확장성)을 보였음
- MAE는 Places에서 이전 최고 성능을 초과 달성. 수십억 개 이미지로 사전학습된 모델들과 경쟁 가능한 수준

Pixels vs. tokens

	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Table 7. **Pixels *vs.* tokens** as the MAE reconstruction target. △ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.

- normalized pixels를 사용하는 것이 통계적으로 유의미하게 더 좋음을 확인할 수 있음.
- pixel 기반 MAE는 BEiT와 달리 복잡한 dVAE 토크나이저가 필요 없음
- MAE에서 tokenizer가 필요하지 않음을 뒷받침하는 실험

6. 결론 (배운점)

💡 연구의 의의 및 한계점, 본인이 생각하는 좋았던/아쉬웠던 점 (배운점)

[연구의 핵심 요약]

- 딥러닝의 핵심 : 단순하면서 확장 가능한 알고리즘 → self-supervised learning 방법이 이에 적합
- NLP 분야와 달리 CV 분야에서는 self-supervised learning이 널리 사용되지 못하고 있었음
- 이 논문에서 효과적인 self-supervised learning 기법인 MAE를 설계함
- 이미지와 언어가 본질적으로 다른 종류의 신호임을 분석하고 이를 바탕으로 구조를 설계함.
    - 이미지는 의미 단위가 아니기 때문에 의미, 구조적 학습을 잘 하지 못한다는 사실을 분석하고, 이미지에 대해서도 의미 단위 학습이 이루어질 수 있도록 하려고 랜덤하게 높은 비율로 패치를 마스킹함.
    - 이를 통해 MAE가 학습한 잠재표현은 의미 있는 구조를 포착하게 됨

[사회적 영향]

- MAE는 훈련 데이터셋에 내재된 통계적 특성 바탕으로 예측하기 때문에 데이터의 bias(편향)을 그대로 반영할 수 있고, 그 중에는 부정적인 사회적 영향을 유발할 수 있는 편향도 포함될 수 있음. 윤리적 고려를 바탕으로 이 기술을 사용해야 함.

[인상깊었던 부분]

- 패치에 마스킹을 이렇게나 많이 하는데 전체적인 구조와 의미 학습이 가능하다는 게 놀랍다.
- 인코더와 디코더 크기를 같지 않게 하는 건 속도 말고 성능에는 어떤 영향을 줄 지 궁금하다.
- NLP의 성능이 좋은 구조를 CV에도 적용하려는 시도가 좋았다. 한 분야의 모델만이 아닌 여러 분야의 모델을 공부해야 하는 이유인 것 같다. 나도 한 모델을 배울 때마다 다른 모델에 어떻게 적용해봐야 할지 고민해봐야겠다.

Swin Transformer

1. Introduction

논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가

논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

논문이 다루는 분야

- Computer Vision
- Transformer 구조를 CV에 적용시키려는 시도
  - CV가 CNN 기반의 길을 걸어온 것과 달리 NLP는 Transformer의 길을 걸어옴
  - NLP에서처럼 CV에서도 Transformer가 backbone으로 사용되게끔 하려는 시도

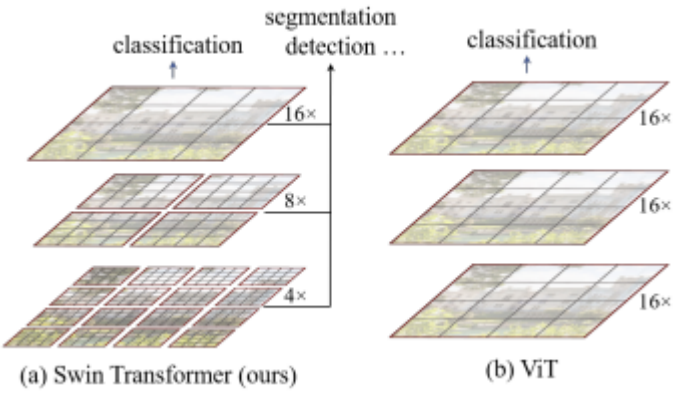
해당 task에서 기존 연구 한계점

[Transformer가 CV에 잘 적용되지 못했던 이유]

1. scale 차이
   - NLP에서는 word token을 사용, 그 크기가 고정적임.
   - visual element는 스케일이 다양해서 고정적 크기를 가정하는 transformer 기반 모델들에서 attention을 적용할 때 문제가 생김.
     - ex> 객체 감지 같은 task는 다양한 크기의 관심 영역 가짐
2. high resolution 차이
   - 시각 이미지의 해상도가 훨씬 높음. 단어토큰 몇십개, 이미지 픽셀은 수천개
   - 기존 transformer기반 모델들은 self-attention의 계산 복잡도가 입력 이미지의 크기에 따라 제곱으로 증가함

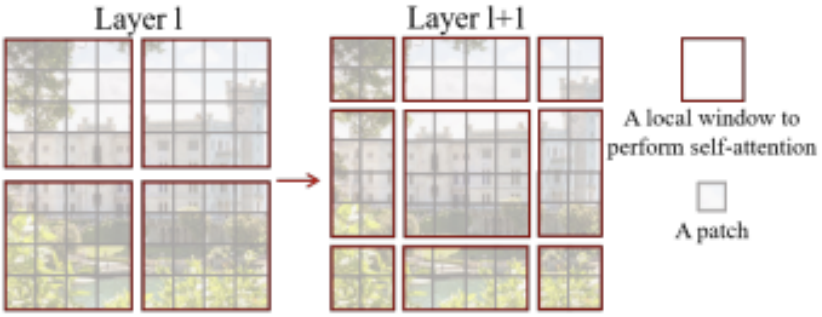
논문의 contributions

- Swin Transformer라는 새로운 범용 Transformer 백본을 제안
  1. 계층적 피처 맵(hierarchical feature map)을 구성



  - 작은 크기 패치(회색)를 시작으로 인접한 패치 병합하며 깊은 계층으로 나아가는 계층적 표현을 구축

2. self-attention의 계산 복잡도를 이미지 크기에 대해 선형적이게 줄임
    - 서로 겹치지 않는 창(window) 단위로 local self-attention을 수행하여 선형 복잡도 달성
    - 창의 개수는 고정되어 있으므로, 복잡도는 이미지 크기에 대해 선형으로 증가
3. **shifted window** self-attention



- 이전 계층에서 정의된 창의 위치를 밀어서 새로운 창을 형성
- 서로 다른 위치의 패치들이 상호작용할 수 있도록 함
- 지역 윈도우 기반 attention의 계산 복잡도 장점을 유지하면서, 전역 모델링 능력을 향상시킴
- 실제 하드웨어에서도 매우 효율적, 다양한 태스크에서 성능 향상을 가져옴
- (sliding window 방식과 달리) shifted window 방식은 낮은 지연 시간을 유지 & 유사한 성능

⇒ 기존 Transformer 기반 구조와 달리, 다양한 해상도에 적응할 수 있는 유연성을 가짐

⇒ 이미지 분류, 객체 검출, 시맨틱 세그멘테이션 등 다양한 태스크에서 강력한 성능을 보임

⇒ 시각과 자연어 처리 전반에 걸쳐 통합된 아키텍처가 존재한다면, 시각 및 언어 신호의 공동 모델링을 가능하게 하고, 양쪽 도메인에서의 지식을 공유함으로써 더 나은 설계를 이끌어낼 수 있을 것이라고 보았음.
