💛

[12주차] SimMIM: a Simple Framework for Masked Image Modeling

0. Abstract

📍 **마스크드 이미지 모델링 프레임워크**

- **Random Masking**: 입력 이미지의 중간 크기(예: 32x32) 패치를 무작위로 마스킹하는 것이 강력한 사전 학습 작업이 됨
- **Pixel Prediction**: 복잡한 설계 없이도 패치 분류 접근 방식만큼 좋은 성능을 내는 것으로 나타났으며, 단순히 원시 픽셀의 RGB 값을 직접 회귀하는 것만으로도 충분함
- **Lightweight Prediction Head**: 선형 레이어만으로도 무거운 예측 헤드만큼 좋은 성능을 낼 수 있음

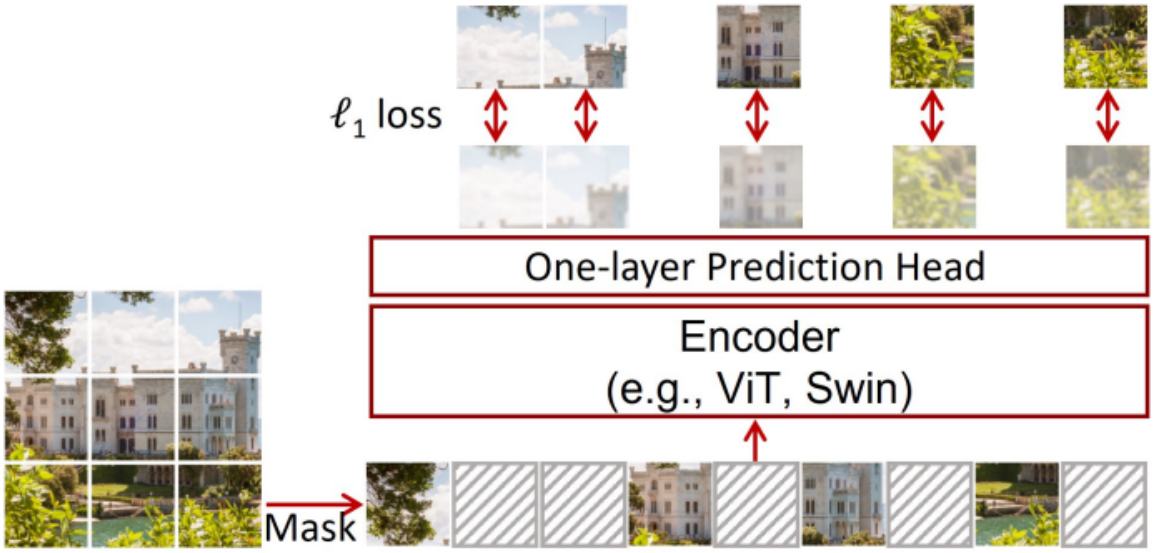
📍 **성능 평가**

- ViT-B 모델로 ImageNet-1K 데이터셋에서 사전 학습 후 fine-tuning 시 83.8%의 top-1 정확도를 달성했으며, 이는 이전 최고 접근 방식보다 0.6% 높음
- 더 큰 SwinV2-H 모델로 ImageNet-1K 데이터셋만 사용해 87.1%의 top-1 정확도를 달성

📍 **데이터 효율성 향상**

SimMIM을 활용해 3B 모델(SwinV2-G)을 40배 적은 레이블 데이터로 학습시켜 4개 비전 벤치마크에서 최신 성과를 달성함

1. Instruction



- 마스크드 신호 모델링의 컴퓨터 비전 적용의 어려움
  - 이미지와 텍스트의 특성 차이로 인해 컴퓨터 비전 분야에 적용하는 것이 어려움
    - 이미지는 강한 지역성을 가지고 있어, 근접한 픽셀들이 높은 상관관계를 가지며, 따라서 의미적 추론보다는 근접 픽셀 복사로 과제를 해결할 수 있음
    - 텍스트 토큰은 고수준 개념인 반면 시각 신호는 저수준 원시 데이터이며, 저수준 신호 예측이 고수준 비전 인식 작업에 유용한지 불확실함
    - 텍스트는 이산적이지만 **시각 신호는 연속적**이기 때문에 이산 토큰 기반 마스크드 언어 모델링 접근법을 연속적 시각 신호에 적용하기 어려움
- **SimMIM**: 간단하지만 강력한 마스크드 이미지 모델링 프레임워크
  - 기존 접근법들은 이러한 어려움을 해결하기 위해 특수한 설계를 도입했음
  - 반면 본 논문에서는 **시각 신호의 특성을 잘 반영**하는 간단한 프레임워크 **SimMIM**을 제안함
    - 입력 이미지 패치에 무작위 마스킹을 적용
    - 마스킹된 영역의 원시 픽셀 값을 ℓ1 loss로 회귀하는 작업을 수행
    - 단순한 선형 레이어만으로도 강력한 성능을 달성할 수 있음
- SimMIM의 성능 및 확장성
  - ViT-B 모델로 ImageNet-1K에서 사전 학습 후 fine-tuning 시 83.8%의 top-1 정확도를 달성
  - 더 큰 SwinV2-H 모델로 ImageNet-1K 데이터셋만 사용해 87.1%의 top-1 정확도를 달성

- SimMIM을 활용해 3B 모델(SwinV2-G)을 40배 적은 레이블 데이터로 학습시켜 4개 비전 벤치마크에서 최신 성과를 달성

2. Related Works

📍 **마스크드 언어 모델링(Masked Language Modeling)**

- 자연어 처리 분야에서 가장 주도적인 자기 지도 학습 접근법
- 문장 내 일부 토큰을 가려내고 이를 예측하는 방식으로 표현 학습을 수행
- 대규모 언어 모델 학습과 다양한 언어 이해/생성 작업에 효과적

📍 **마스크드 이미지 모델링(Masked Image Modeling)**

- 언어 모델링과 병행하여 발전해왔지만 주류 위치에 있지 않았음
- 초기 접근법인 Context Encoder, CPC 등이 있었으나 최근까지 주목받지 못함
- 최근 Vision Transformer 모델에서 이 접근법이 주목받기 시작
- iGPT, ViT, BEiT 등이 픽셀 클러스터링, 평균 색상 예측, 토큰화 등의 특수 설계를 도입

📍 **SimMIM**

- 본 논문의 접근법
- 위 접근법들과 달리 매우 간단한 프레임워크를 제안
- 이미지 패치에 무작위 마스킹을 적용하고 $\ell 1$ loss로 마스킹된 영역의 픽셀 값을 회귀
- 이전 접근법과 유사하거나 더 나은 성능을 보임

<그 외>

리콘스트럭션 기반 방법: 원본 신호 복원을 목표로 함 (본 접근법과 다른 철학)

이미지 인페인팅: 마스크드 이미지 모델링과 연관된 고전적 컴퓨터 비전 문제

압축 센싱: 대부분의 데이터를 버려도 인지적 손실이 크지 않다는 이론적 근거 제공

기타 자기 지도 학습 접근법: 색상 예측, 퍼즐 풀이 등 다양한 프리텍스트 작업 존재

3. Approach

3.1 A Masked Image Modeling Framework

1️⃣ **마스킹 전략**

- 입력 이미지에서 어떤 영역을 가릴지 결정하고, 실제로 마스킹을 수행
- 마스킹된 이미지가 모델의 입력으로 사용됨

2️⃣ **인코더 아키텍처**

- **마스킹된 이미지의 잠재 특징 표현**을 추출함
- 다양한 비전 작업에 활용될 수 있는 것이 기대됨
- 본 논문에서는 ViT와 Swin Transformer, 두 가지 비전 Transformer 아키텍처를 사용함

3️⃣ **예측 헤드**

- **잠재 특징 표현**을 이용하여 **가려진 영역의 원래 신호를 예측**함

4️⃣ **예측 대상**

- 가려진 영역의 원래 신호 형태를 정의
- **원 픽셀 값 또는 이를 변환한 형태**를 예측 가능
- 손실 함수로는 교차 엔트로피 분류 손실이나 $\ell 1$, $\ell 2$ 회귀 손실 등을 사용할 수 있음

3.2 Masking Strategy

Figure 2. Illustration of masking area generated by different masking strategies using a same mask ratio of 0.6: square masking [38], block-wise masking [1] apply on 16-sized patches, and our simple random masking strategy on different patch sizes (e.g., 4, 8, 16 and 32).

1️⃣ **패치 정렬 랜덤 마스킹**

- **이미지를 패치 단위로 처리**하므로, **패치 단위**로 마스킹을 수행
- 패치는 완전히 보이거나 완전히 가려짐
- **Swin Transformer**의 경우 4x4~32×32 크기의 패치를 고려하고, 기본적으로 32x32 패치를 사용
- **ViT**의 경우 기본적으로 32x32 크기의 패치를 사용

2️⃣ **그 외 마스킹 전략**

- **중앙 영역 마스킹**: 이전 연구에서 제안된 방식을 응용하여, 중앙 영역을 랜덤하게 이동시키며 마스킹

- **블록 단위 마스킹**: BEiT에서 제안된 복잡한 블록 단위 마스킹 전략을 시도합니다. 16x16 및 32x32 크기의 패치를 사용

3.3 Prediction Head

- **Prediction Head**는 임의의 형태와 용량을 가질 수 있음
- 단 Encoder의 출력과 일치하고 Prediction Target을 달성할 수 있어야 함
- 이전 연구에서는 무거운 Prediction Head(Decoder)를 사용
- 본 논문에서는 **선형 레이어와 같이 매우 가벼운 Prediction Head도 충분히 효과적임**을 보여줌
- 2-layer MLP, 역 Swin-T, 역 Swin-B와 같은 더 무거운 Head도 시도

3.4 Prediction Targets

1️⃣ Raw Pixel Value Regression

- 픽셀 값은 연속적인 색상 공간이므로, **마스킹된 영역의 원래 픽셀 값을 회귀**하는 것이 직관적
- 대부분의 비전 아키텍처는 입력 이미지보다 낮은 해상도의 특징 맵을 출력함
- 각 특징 벡터를 원래 해상도로 매핑하여 해당 픽셀 값을 예측하도록 함
- 이때 $\ell 1$ loss를 사용하여 **마스킹된 픽셀 값과의 차이를 최소화**
- 원본 이미지를 **다운샘플링**하여 낮은 해상도 타겟으로 실험 진행

2️⃣ 그 외 Prediction Targets

- 이전 연구에서는 마스킹된 신호를 클러스터나 클래스로 변환하고 분류 작업을 수행
- **색상 클러스터링**: iGPT에서 사용한 512개 클러스터 중심을 사용
- **비전 토큰화**: BEiT에서 사용한 이산 VAE 네트워크를 통해 토큰 ID를 예측함
- **채널별 bin 색상 이산화**: R, G, B 채널을 각각 이산화된 bin으로 분류

3.5 Evaluation protocols

- ImageNet-1K 이미지 분류 fine-tuning 성능을 주요 평가 지표로 사용
  - 실제 응용 시나리오에 더 유용한 지표입니다.
- 이전 연구에서 사용했던 선형 프로빙 성능도 함께 보고했으나 주요 목표는 ❌
- 본 연구의 주된 목적은 다운스트림 작업을 잘 보완할 수 있는 표현을 학습하는 것

4. Experiments

4.1 Ablation study

📍 Settings

- Backbone 및 입력 크기
  - Swin-B 모델을 기본 backbone으로 사용
  - 입력 이미지 크기는 192x192로 설정했고, 이에 맞춰 Swin 모델의 윈도우 크기를 6으로 조정함
  - ImageNet-1K 데이터셋을 사용했으며, 사전 학습과 fine-tuning에 모두 활용
- 사전 학습 설정
  - AdamW 옵티마이저와 코사인 학습률 스케줄러를 사용
  - 100 epoch 동안 학습 진행
- SimMIM 기본 설정
  - 랜덤 마스킹, 32x32 패치 크기, 60% 마스크 비율
  - 선형 예측 헤드, 192x192 타겟 이미지 크기
  - $\ell 1$ loss 사용
- Fine-tuning 설정
  - AdamW 옵티마이저, 100 epoch, 코사인 학습률 스케줄러
  - 10 epoch 동안 warm-up
  - 데이터 증강: RandAug, Mixup, Cutmix, 라벨 스무딩, 랜덤 Erasing

📍 Masking Strategy

- 마스킹 전략이 성능에 미치는 영향
  - 다양한 마스킹 전략과 마스킹 비율에 따른 fine-tuning 정확도를 비교
  - 단순한 랜덤 마스킹 전략이 83.0%의 최고 정확도 달성
    - 이전 연구에서 제안된 블록 단위 마스킹보다 0.3% 높은 성능
- 마스킹 패치 크기와 마스킹 비율
  - 32×32 크기의 큰 마스킹 패치를 사용하면 10-70% 범위의 다양한 마스킹 비율에서 안정적으로 좋은 성능을 보임

        - 큰 마스킹 패치의 중심 픽셀은 가시 픽셀들과 충분히 멀리 떨어져 있어, 네트워크가 장거리 관계를 학습하도록 유도
    - 작은 패치 크기(4, 8, 16)에서는 마스킹 비율을 높이면(0.4 → 0.8) 성능이 점진적으로 향상되었음
        - 하지만 32x32 패치 크기에 비해 전반적인 정확도는 ⬇️
    - 패치 크기를 64x64로 더 늘리면 오히려 성능이 저하되었음
        - 예측 거리가 너무 커져 학습이 어려워진 것으로 보임
- AvgDist 지표
    - 본 논문에서 마스킹된 픽셀과 가장 가까운 가시 픽셀 간 유클리드 거리의 평균인 AvgDist 지표를 제안
    - AvgDist는 마스킹 비율이 증가할수록 전반적으로 증가
        - 작은 패치 크기(4, 8)에서는 AvgDist가 낮고 천천히 증가
        - 큰 패치 크기(64)에서는 낮은 마스킹 비율(10%)에서도 AvgDist가 높은 것으로 나타남
        - 블록 단위 마스킹과 정사각형 마스킹은 64x64 패치와 유사한 높은 AvgDist를 보임
    - 정확도와 AvgDist의 관계는 ridge 형태를 보임
        - 정확도가 높은 지점은 AvgDist 범위 10-20 사이에 분포
        - AvgDist가 너무 작거나 크면 성능이 저하됨
- 최종 설정
    - 안정적인 성능을 위해 32x32 패치 크기, 60% 마스킹 비율을 기본 설정으로 사용
    - 언어 모델과 달리 이미지에서는 더 높은 마스킹 비율이 효과적인 것으로 나타남
    - 두 모달리티의 정보 중복성 차이 때문으로 추정

📍 Prediction Head

- 다양한 예측 헤드(linear layer, 2-layer MLP, inverse Swin-T, inverse Swin-B)의 효과를 비교
- 일반적으로 더 복잡한 헤드가 약간 낮은 loss를 보임
    - 예를 들어, inverse Swin-B는 0.3722의 loss, 선형 레이어는 0.3743의 loss를 보임

- 그러나 ImageNet-1K 태스크에서의 전이 성능은 오히려 더 낮은 것으로 나타남
  - 이는 더 강력한 inpainting 능력이 반드시 더 나은 다운스트림 성능으로 이어지지 않음을 보여줌
  - 예측 헤드의 복잡도가 높으면 다운스트림 태스크에서 활용되지 않는 용량이 낭비될 수 있기 때문
- 복잡한 헤드는 훈련 비용도 더 높음
  - inverse Swin-B 헤드는 선형 레이어 대비 2.3배 더 높은 비용이 들었음

📍 Prediction Resolution

- 다양한 예측 해상도(62 ~ 1922)의 효과를 비교
- 대부분의 해상도(122 ~ 1922)에서 성능이 유사함
- 다만 62의 낮은 해상도에서는 성능이 떨어졌음
  - 이는 너무 많은 정보가 손실되어 이미지 분류 태스크에 필요한 정보 수준을 충족하지 못했기 때문으로 보임
- 저자들은 실험에서 1922 해상도를 기본값으로 사용했음
  - 이 해상도에서 최고 수준의 전이 성능을 보이면서도 계산 비용이 크지 않기 때문

📍 Prediction Target

- 다양한 예측 타겟(ℓ1 loss, smooth-ℓ1 loss, ℓ2 loss, 색상 클러스터링 기반 클래스, 토큰화 기반 클래스, 채널별 균등 이진화)의 효과를 비교
- 주요 발견은 다음과 같음
  - ℓ1, smooth-ℓ1, ℓ2 loss는 유사한 성능을 보임
  - 색상 클러스터링 기반 클래스나 토큰화 기반 클래스는 제안한 방식보다 약간 낮은 성능을 보임
  - 채널별 균등 이진화 방식은 ℓ1 loss와 유사한 성능을 보였지만, 이진화 구간 수 튜닝이 필요
- 이를 통해 마스크된 이미지 모델링의 타겟을 마스크드 언어 모델링과 동일한 분류 기반으로 맞출 필요는 없다는 것을 알 수 있음
- 시각 신호의 고유한 특성에 맞추어 접근하는 것이 좋음
- **예측 vs 재구성**

Scope to predict	Top-1 acc (%)
masked area	82.8
full image	81.7

Table 4. Ablation on different performing areas of prediction loss. If the loss is computed at masked area, it performs a pure prediction task. If it is computed on the whole image (both masked & unmasked areas), it performs a joint prediction and reconstruction task.

- 오토인코더와 마스크드 이미지 모델링은 모두 원본 신호를 복원하지만, 접근 방식이 다름
  - 오토인코더: 가시 신호의 재구성에 초점을 맞춤
  - 마스크드 이미지 모델링: 비가시 신호의 예측에 초점을 맞춤
- 실험 결과, 마스크된 영역만 예측하는 방식이 전체 이미지를 재구성하는 방식보다 성능이 더 좋은 것으로 나타남(82.8% vs 81.7%).

➡️ 두 작업이 내부 메커니즘에서 근본적으로 다르며, 예측 작업이 더 효과적인 표현 학습 접근법일 수 있음을 시사함

4.2 Comparison to Previous Approaches on ViT-B

- 이전 연구들이 ViT 아키텍처에서 실험을 수행했기 때문에, 공정한 비교를 위해 ViT-B 아키텍처를 사용
- 사전 학습 시 800 epoch, 코사인 학습률 스케줄러, 20 epoch 선형 워밍업을 사용
- 미세 조정 시 레이어별 학습률 감쇄 0.65를 사용
- 선형 프로빙 시 ViT-B의 중간 레이어를 선택했고, 100 epoch 학습에 5 epoch 선형 워밍업을 사용
- 제안한 SimMIM 방식은 미세 조정 시 83.8% top-1 정확도를 달성했는데, 이는 이전 최고 접근법 대비 0.6% 높은 수치
- SimMIM은 학습 효율성이 2.0배, 1.8배, 약 4.0배, 1.5배 더 높은 것으로 나타남
- 선형 프로빙 정확도도 보고했지만, 주 초점은 미세 조정 성능 향상임

4.3. Scaling Experiments with Swin Transformer

- Swin Transformer의 다양한 모델 크기(Swin-B, Swin-L, SwinV2-H, SwinV2-G)를 실험함
- 사전 학습 시 192^2 입력 해상도, 800 epoch, 4e-4 초기 학습률, 7/8 epoch 후 0.1 감쇄를 사용
- 미세 조정 시 224^2 해상도, 100 epoch(SwinV2-H는 50 epoch), 레이어별 학습률 감쇄 0.8, 0.75, 0.7을 사용
- SimMIM 사전 학습을 거친 모든 Swin 모델이 지도 학습 모델보다 유의미하게 높은 정확도를 달성
- SwinV2-H 모델에 5122 해상도를 적용하면 87.1% top-1 정확도를 달성했는데, 이는 ImageNet-1K 데이터만 사용한 방법 중 최고 성능
- 이전 연구들은 거대 JFT-3B 데이터셋을 사용했지만, SimMIM은 40배 작은 데이터로도 SwinV2-G 3B 모델을 강력하게 학습시킬 수 있었음

4.4 Visualization

📍 **학습된 능력**

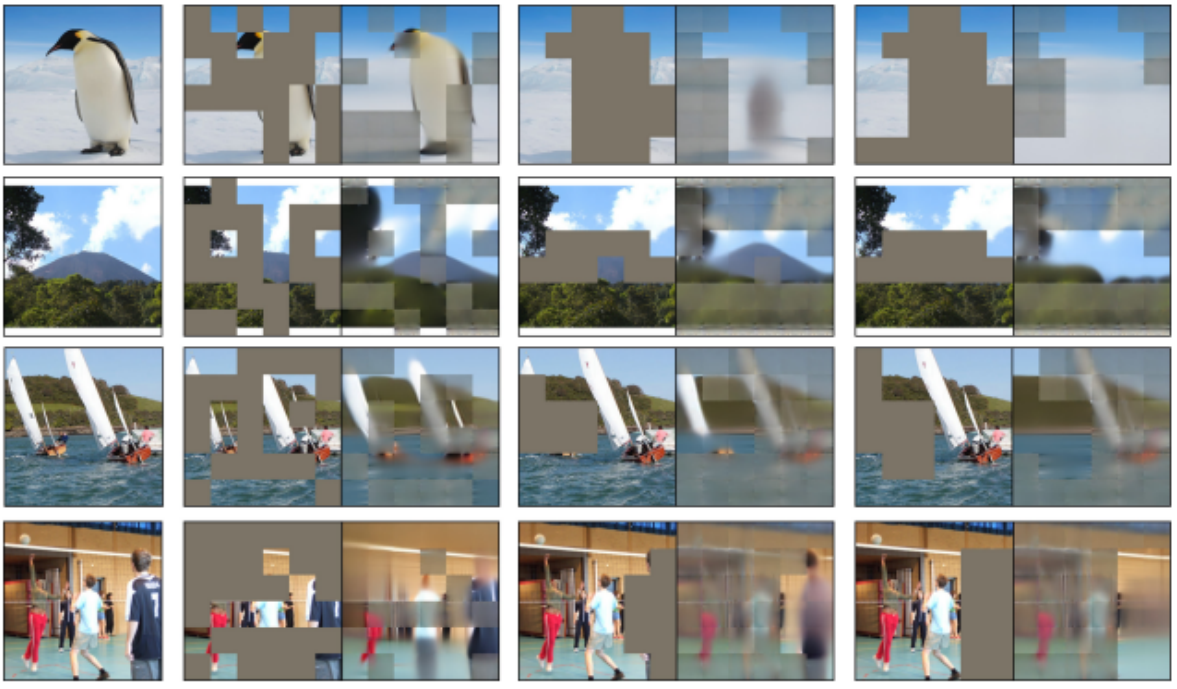


Figure 4. Recovered images using three different mask types (from left to right): random masking, masking most parts of a major object, and masking the full major object.

- Fig4는 사람이 직접 설계한 마스크를 사용하여 복원된 이미지를 보여줌

- 랜덤 마스크로 주요 객체의 중간 정도를 가린 경우, 마스킹된 부분의 형태와 질감을 잘 복원할 수 있음
- 주요 객체의 대부분(90% 이상)을 가린 경우에도 모델은 미미한 단서로 객체의 존재를 예측 가능함
- 객체가 완전히 가려진 경우, 모델은 배경 텍스처로 마스킹된 영역을 채워넣음 → 모델이 객체에 대한 강력한 추론 능력을 학습했음을 파악 가능
    - 단순히 이미지 ID를 기억하거나 주변 픽셀을 복사하는 것 ❌

📍 예측 vs 복원

Figure 5. Recovered images by two different losses of predicting only the masked area or reconstructing all image area, respectively. For each batch, images from left to right are raw image, masked image, prediction of *masked patches only*, and reconstruction of *all patches*, respectively.

- 위의 Table 4에서 마스크된 예측 작업만 수행한 경우가 마스크된 예측과 가시 신호 복원을 함께 수행한 경우보다 성능이 크게 좋은 것으로 나타남
- Fig5에서도 후자의 접근법이 더 좋은 복원 결과를 보이지만, 가시 영역 복원에 모델 용량이 낭비되어 미세 조정 성능이 떨어지는 것으로 볼 수 있음

📍 마스크 패치 크기의 영향

Figure 6. An example of recovered image using masked patch sizes of 4, 8, 16, 32 and 64, and a fixed masked ratio of 0.6.

- Fig6은 마스킹 비율 0.6에서 마스크 패치 크기를 변경했을 때의 복원 결과를 보여줌
- 패치 크기가 작을수록 세부 사항을 더 잘 복원할 수 있지만, 전이 학습 성능은 떨어짐
- 패치 크기가 작으면 주변 픽셀이나 텍스처만으로도 예측 작업을 쉽게 수행할 수 있기 때문이라고 볼 수 있음

5. Conclusion

📍 SimMIM 프레임워크의 주요 특징

1️⃣ 랜덤 마스킹 전략: 중간 크기의 패치를 무작위로 마스킹

2️⃣ 직접 회귀 예측: RGB 픽셀 값을 직접 회귀하여 예측

3️⃣ 간단한 예측 헤드: 선형 레이어만으로도 충분한 예측 헤드를 사용함

👉 단순한 설계에도 불구하고 SimMIM은 강력한 성능을 보여줌

👉 본 논문의 저자들은 이처럼 간단한 프레임워크와 강력한 결과가 향후 이 분야의 연구를 촉진하고, AI 분야 간 심도 있는 상호작용을 장려할 수 있기를 희망하고 있음

논문에 대한 의견 및 의문점(꼭지)

➡️ 본 논문에서는 마스킹 기법을 주로 사용하고 있는 모습을 볼 수 있는데, 이미지 분류 외에도 마스크된 이미지 모델링을 다른 컴퓨터 비전 작업에 적용할 수 있는 방법을 모색해보면 좋을 것 같음. 특히 마스크된 이미지 모델링을 자연어 처리, 음성 인식 등 다른 AI 분야로 확장하는 방법을 고려해볼 수 있을 것이라 생각함