visual perception

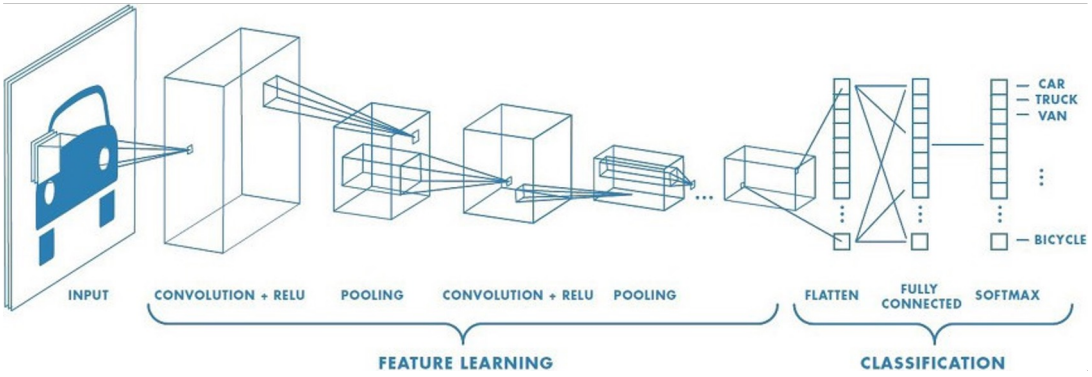
10.01 이현수 최주연

Image Classification



머핀? 강아지?

CNN


CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION

CNN

- **귀납적 편향이 강함**
  - 지역성, 평행 이동 불변성 같은 특성이 아키텍처에 내장돼 있음
  - 작은 데이터셋에서는 장점이지만, 대규모 데이터셋에서는 오히려 모델이 학습 가능한 표현의 다양성을 제한하여 유연성 부족

- **전역적 관계 학습이 어려움**
  - 주로 지역적 패턴을 학습함
  - 이미지 전체의 장거리 의존성을 직접적으로 모델링하기 어려움

- 특화된 설계와 연산 필요
  - 컨볼루션, 풀링 같은 연산은 이미지 특화 연산이라 범용적이지 않음

Transformer

- **Self Attention**을 통해 입력 전체를 동시에 보.
- 전역적인 관계를 잘 학습할 수 있음

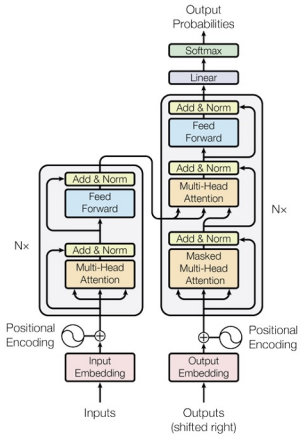
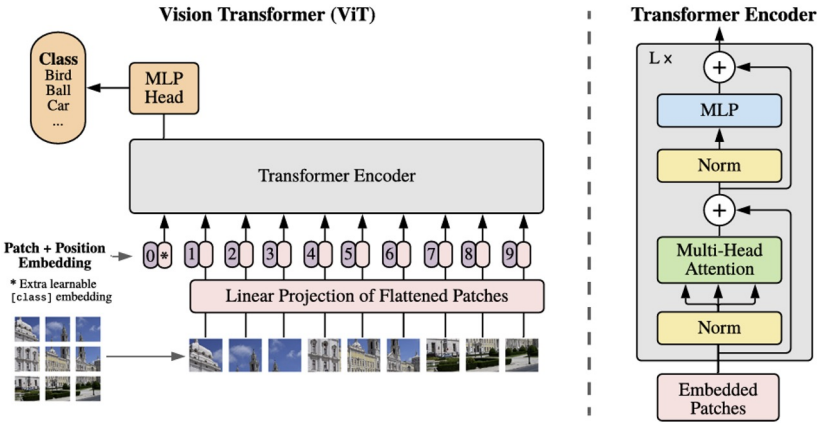


Figure 1: The Transformer - model architecture.

Vision Transformer

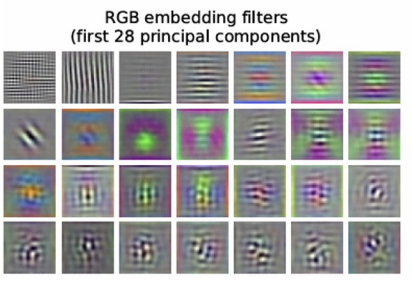
AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

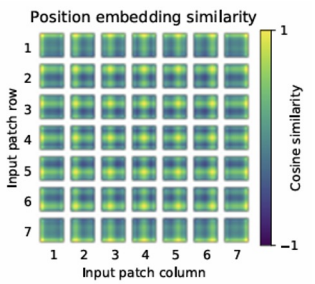
Vision Transformer

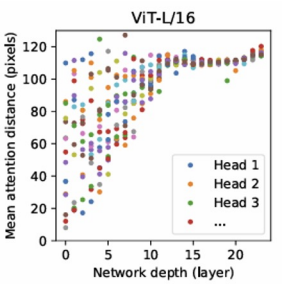


Vision Transformer

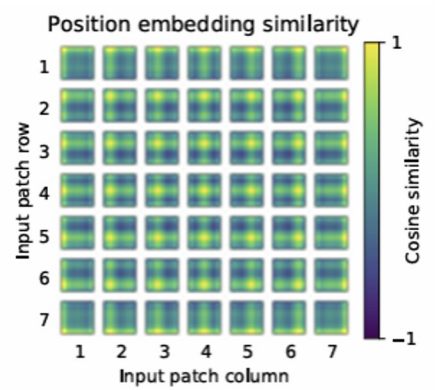
Self Supervision







Vision Transformer

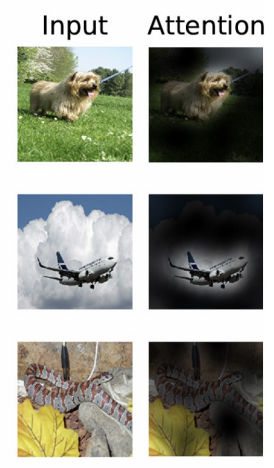


Inspecting Vision Transformer

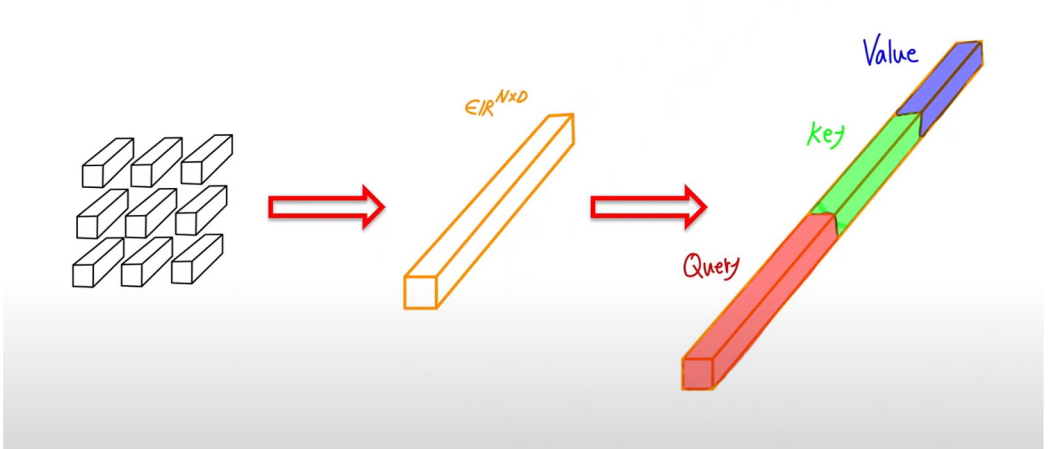
Patch Embedding

Position Embedding

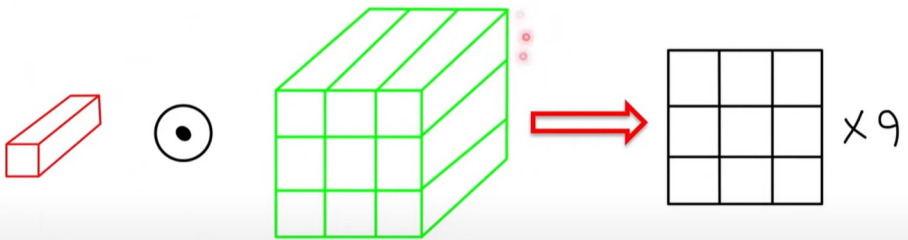
Self-Attention 시각화



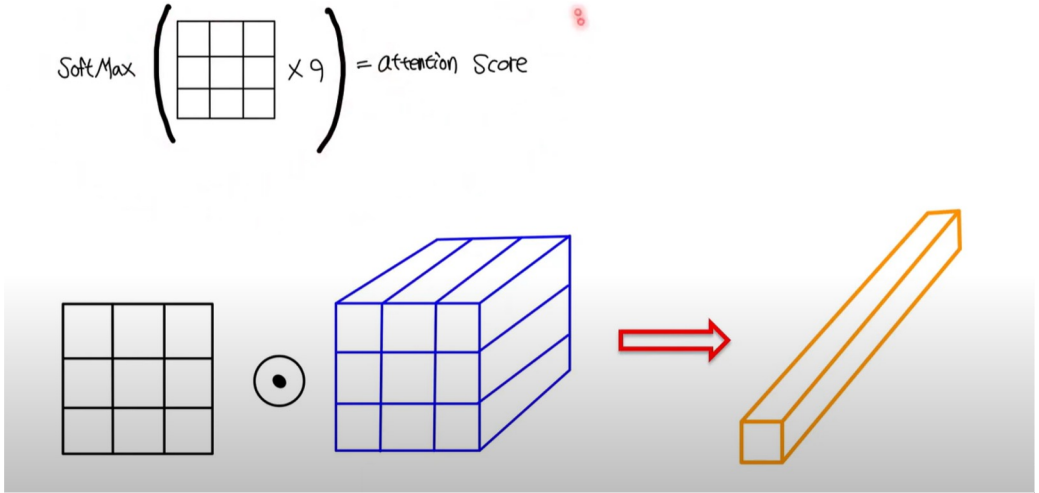
Vision Transformer



Vision Transformer



Vision Transformer



Vision Transformer

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots ; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z'}_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \dots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \qquad \ell = 1 \dots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

Vision Transformer - setting

- Dataset
  - ImageNet (1.3M)
  - ImageNet-21k (14M)
  - JFT-300M (300M, Google 내부 데이터셋)
  - 다운스트림 전이 학습: CIFAR, VTAB, Pets, Flowerst
- Model
  - Base, Large, Huge

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

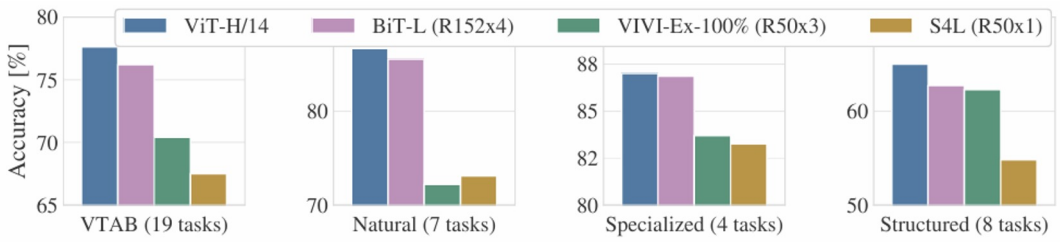
Vision Transformer - setting

- Model Variants
  - ViT-B, ViT-L, ViT-H → BERT 모델 크기 구분 방식 차용
  - 입력 패치 크기: 예) ViT-L/16 = Large 모델, 패치 크기 16×16
  - 패치 크기가 작아질수록 입력 시퀀스 길이가 길어져 연산량 ↑

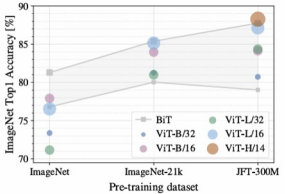
Vision Transformer - setting

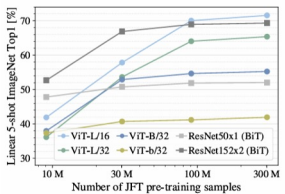
- Training & Fine-tuning
  - Optimizer: Adam (사전학습), SGD (파인튜닝)
  - Batch size: 4096 (pre-training), 512 (fine-tuning)
  - Weight decay: 0.1, learning rate warm-up/decay 사용
  - Fine-tuning: 더 높은 해상도(예: 224→384)로 수행
- Metrics
  - Downstream task 성능 → fine-tuning 정확도 기준
  - Few-shot accuracy도 함께 보고, on-the-fly 평가에 활용

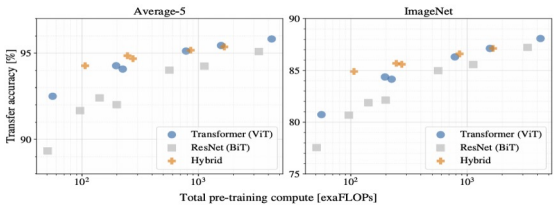
Vision Transformer - experiment



Vision Transformer - experiment







Self-Supervision

ViT-B/16 모델이 ImageNet에서 79.9% 정확도

Appendix A. Multihead Self-Attention

Standard **qkv** self-attention (SA, Vaswani et al. (2017)) is a popular building block for neural architectures. For each element in an input sequence $\mathbf{z} \in \mathbb{R}^{N \times D}$, we compute a weighted sum over all values $\mathbf{v}$ in the sequence. The attention weights A_{ij} are based on the pairwise similarity between two elements of the sequence and their respective query $\mathbf{q}^i$ and key $\mathbf{k}^j$ representations.

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv} \qquad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h}, \tag{5}$$

$$A = \mathrm{softmax}\left(\mathbf{q}\mathbf{k}^\top / \sqrt{D_h}\right) \qquad A \in \mathbb{R}^{N \times N}, \tag{6}$$

$$\mathrm{SA}(\mathbf{z}) = A\mathbf{v}\,. \tag{7}$$

Multihead self-attention (MSA) is an extension of SA in which we run k self-attention operations, called "heads", in parallel, and project their concatenated outputs. To keep compute and number of parameters constant when changing k, D_h (Eq. 5) is typically set to D/k.

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(z); \mathrm{SA}_2(z); \cdots; \mathrm{SA}_k(z)]\,\mathbf{U}_{msa} \qquad \mathbf{U}_{msa} \in \mathbb{R}^{k \cdot D_h \times D} \tag{8}$$



DINOv3 🦖🦖🦖

20230062 이현수

What is DINOv3?

: self-supervised learning for vision at unprecedented scale

: A high-performance vision backbone that is universally applicable to large-scale image data.

https://www.youtube.com/watch?v=-eOYWK6m3i8

1. Scaling both dataset and model size by careful data preparation, design,, and optimization
2. Gram anchoring
3. Post-hoc strategies

background

- **Knowledge destillation**

: teacher model의 knowledge를 student model에 전달하는 기술

7B parameter model ViT-S, ViT-B

🦕 **student-teacher network**

: 딥러닝 모델에서 지식을 전달하는데 사용하는 방법

🦕 *teacher network*

- training 과정에서 weigh가 고정되거나, student network의 weight의 EMA(Exponential Moving Average)로 업데이트
- Student network보다 안정적

🦕 *student network*

- Training 중 weight update

background

• feature map

: 이미지에 있는 시각적 pattern을 탐지하기 위해 kernel을 적용해서 만들어지는 2차원 배열

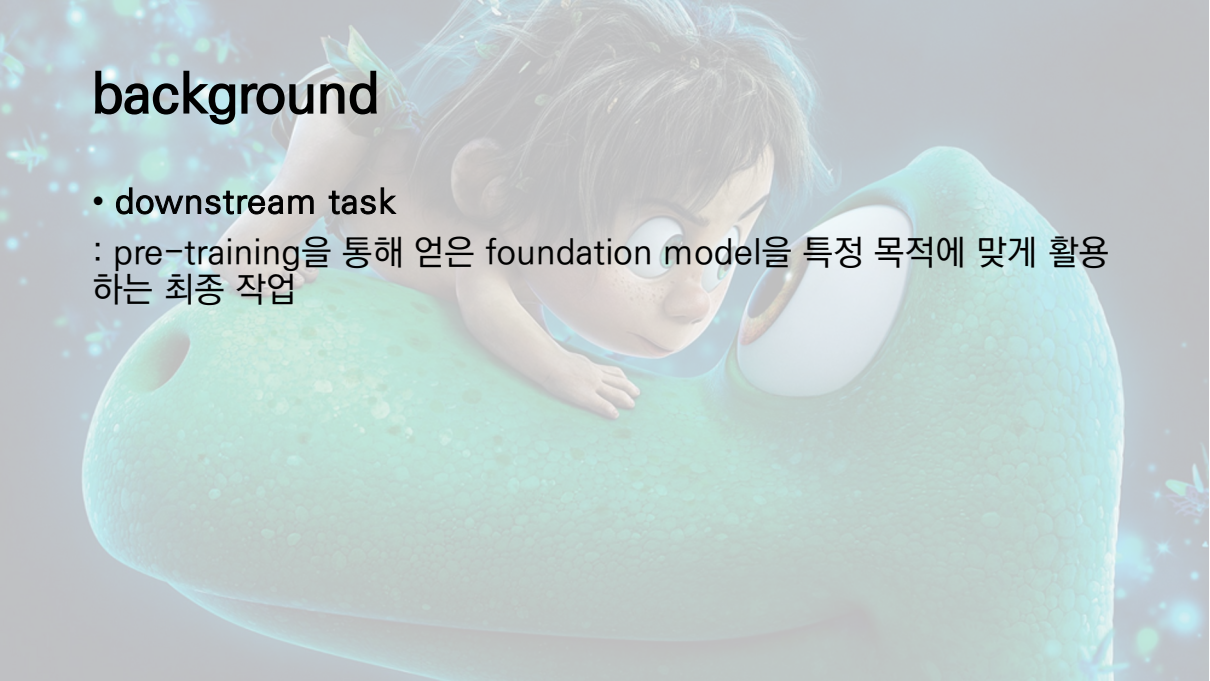
🦖 dense feature map

: 특정 부분의 patter을 탐지

background

- downstream task

: pre-training을 통해 얻은 foundation model을 특정 목적에 맞게 활용하는 최종 작업



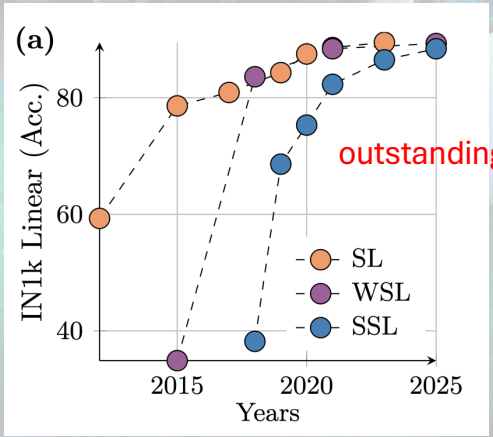
background

- **model scaling**

: model의 capacity(e.g. # of parameters), training data 양, computing 자원 등을 증가

e.g. DINOv2 1M

SSL(Self-Supervised Learning)



SSL의 기존 문제점

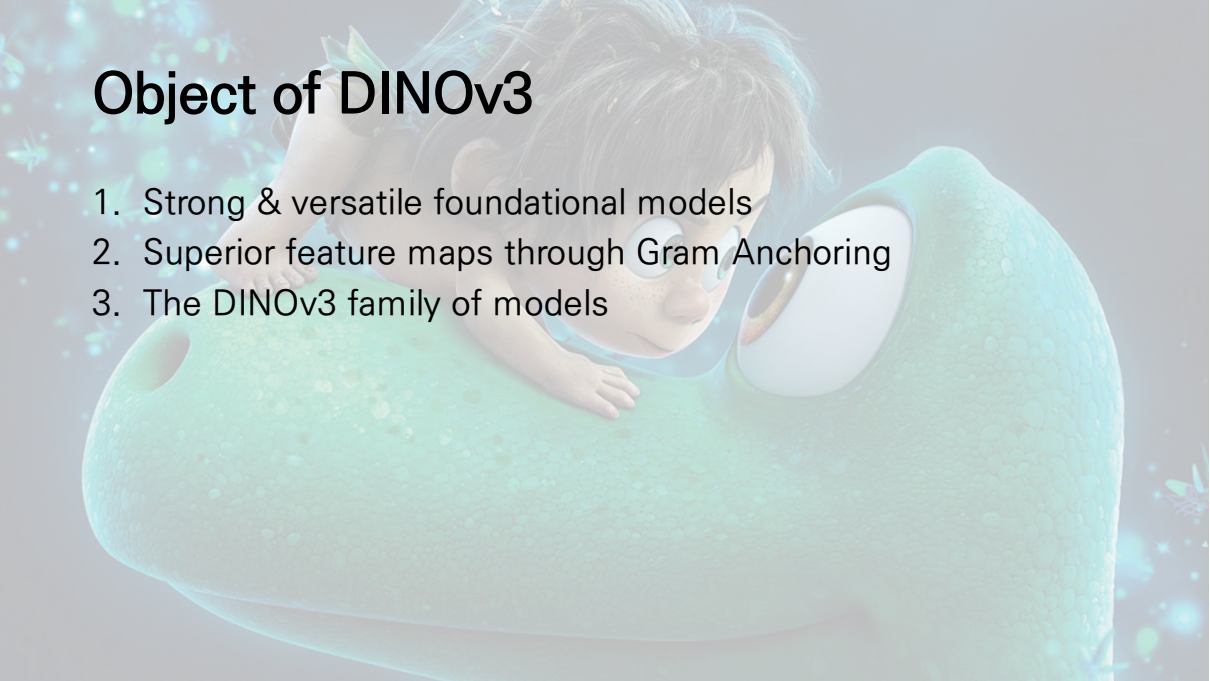
: scaling에서 어려움을 겪음.

1. Unlabeled data에서 useful data를 collect하는 것이 불분명함
2. Dense feature map이 점진적으로 저하되며 patch 간 코사인 유사도 맵이 노이즈가 많아지고, 부정확해짐
3. Feature performace가 학습을 거듭할 수록 감소함

Advanced SSL training at scale

Object of DINOv3

1. Strong & versatile foundational models
2. Superior feature maps through Gram Anchoring
3. The DINOv3 family of models



Strong & versatile foundational models

GOAL

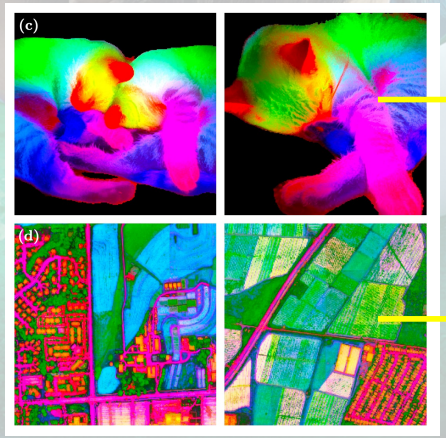
- training data와 model을 scaling하여 high level of versatility를 얻는 것
- 이를 통해

1. weight가 frozen된 상태에서 좋은 performance를 내기
2. 다양한 scientific application에 적용하기

를 기대할 수 있음.

- 즉, 별도의 tuning 없이 성능을 높이고, scientific application을 하겠다는 것

Strong & versatile foundational models



Strong & versatile foundational models

🦕 스포하기 …

- 세계자원연구소: 위성 이미지로 산림 훼손과 토지 이용 변화를 감지
- NASA JPL: 화성 탐사
- ORAKI Oncology: 장기 유사체 이미지로 pre training 시켜, 암 치료에 대한 암환자의 반응 예측

Superior Feature Maps Through Gram Anchoring

- dense feature map의 significant improvement
- High-level semantic task에서 performance를 높이는 것이 목표
- dense representation + global representation 방식이 방대한 양의 이미지 데이터를 train할 때는 optimize하기 어려움

(high-level understanding이 dense feature map의 quality와 conflict)

Superior Feature Maps Through Gram Anchoring

스포하기 …

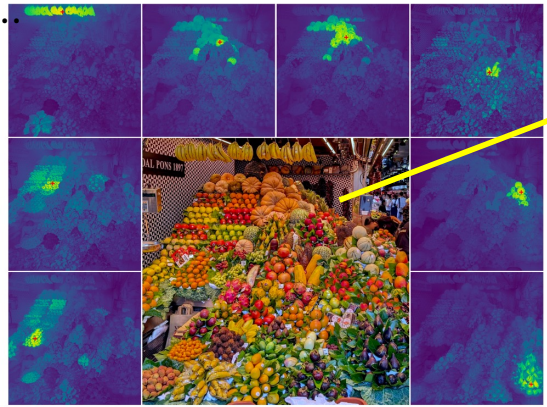


Figure 3: High-resolution dense features. We visualize the cosine similarity maps obtained with DINOv3 output features between the patches marked with a red cross and all other patches. Input image at 4096×4096. *Please zoom in*, do you agree with DINOv3?

4096 x 4096

patch와 image의 다른 부분의 cosine similarity 계산

The DINOv3 Family of Models

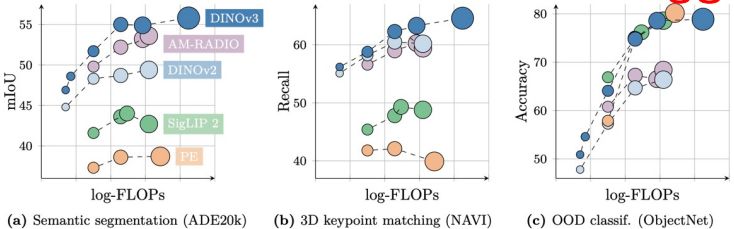
- Dense feature map의 degradation을 해결함으로써 scaling에 대한 제약이 줄어듦 → SSL을 이용한 더 큰 model의 improvement
- ViT + ConvNeXt-based architectures



The DINOv3 Family of Models

🦕 스포하기 …

성능 최고!



(a) Semantic segmentation (ADE20k) **(b)** 3D keypoint matching (NAVI) **(c)** OOD classif. (ObjectNet)

Figure 2: Performance of the DINOv3 family of models, compared to other families of self- or weakly-supervised models, on different benchmarks. DINOv3 significantly surpasses others on dense benchmarks, including models that leverage mask annotation priors such as AM-RADIO (Heinrich et al., 2025).

Training at scale without supervision

- Data preparation
- Large-scale training with self-supervision

- Intend similar paradigm shift for computer vision



1 Data Preparation

- Data scaling – 양적 확장 + data 선별
- Training data의 size만을 키우는 것은 model의 performance를 향상시키는 데 도움이 안 됨 → careful **data curation** pipline!
- DINOv3를 위해 2개의 complementary approach를 선택

Data Preparation – collection & curation

- Automatic curation method: hierarchical k-means
- Retrieval-based curation system

→ 모두 선행 연구 기법/ platform level에서 어느정도 검증

+) ImageNet과 같이 available computer vision data 사용

Data Preparation – sampling

- 앞서 수집한 data 중 어떤 것을 골라 model에 학습시킬지 결정

1. *Homogeneous batches*: 하나의 dataset에서만 sample을 가져와 batch를 구성
2. Heterogeneous batches: 여러 dataset에서 sample을 가져와 batch를 구성

cf) batch: 전체 training data set을 여러 개의 작은 그룹으로 나눈 것 (for multi gpu)

Data Preparation – ablation

- Ablation?

: model의 다양한 data 조합이 성능에 미치는 영향을 평가하기 위해, 특정 component를 체계적으로 제거하고 실험하는 과정(like 변인통제)

Dataset	IN1k k-NN	IN1k Linear	ObjectNet	iNaturalist 2021	Paris Retrieval
Raw	80.1	84.8	70.3	70.1	63.3
Clustering	79.4	85.4	72.3	81.3	85.2
Retrieval	84.0	86.7	70.7	86.0	82.7
LVD-1689M (ours)	**84.6**	**87.2**	**72.8**	**87.0**	**85.9**

better performance than raw data

Large-scale training with self-supervision

- Model의 크기 자체를 키우는 것이 performance를 높이는 데 도움
- 훈련 pipeline 개선

1. Learning objective
2. Updated model architecture
3. optimization

Large-scale training with self-supervision – learning objective

이미지 수준 목표

특징이 공간에 균일하게 퍼지게끔

$$\mathcal{L}_{\text{Pre}} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + 0.1 * \mathcal{L}_{\text{DKoleo.}}$$

패치 수준 목표

Large-scale training with self-supervision – updated model architecture

Teacher model	DINOv2	DINOv3
Backbone	ViT-giant	ViT-7B
#Params	1.1B	6.7B
#Blocks	40	40
Patch Size	14	16
Pos. Embeddings	Learnable	RoPE
Registers	4	4
Embed. Dim.	1536	4096
FFN Type	SwiGLU	SwiGLU
FFN Hidden Dim.	4096	8192
Attn. Heads	24	32
Attn. Heads Dim.	64	128
DINO Head MLP	4096-4096-256	8192-8192-512
DINO Prototypes	128k	256k
iBOT Head MLP	4096-4096-256	8192-8192-384
iBOT Prototypes	128k	96k

Large-scale training with self-supervision – optimization

- right optimization horizon을 추측하는 것은 매우 어렵

← model capacity와 training data complexity는 사전에 평가하기 어렵기 때문

- 이를 극복하기 위해 parameter scheduling을 빼고

constant한 (1) learning rate, (2) weight decay, (3) teacher EMA momentum으로 training

Large-scale training with self-supervision – optimization

- 장점

1. Downstream performance가 좋아질 때까지 training할 수 있음
2. # of optimization hyperparameter ↓ – 고르기 더 쉬워짐

constant한 (1) learning rate, (2) weight decay, (3) teacher EMA momentum으로 training

Large-scale training with self-supervision – optimization

- AdamW
- total batch size = 4096 & 256 GPU
- Multi-crop strategy

: model이 다양한 크기와 해상도의 이미지로부터 특징을 학습할 수 있게 하는 data augmentation 기술

: teacher – 대형 crop, student – 대형/소형 crop

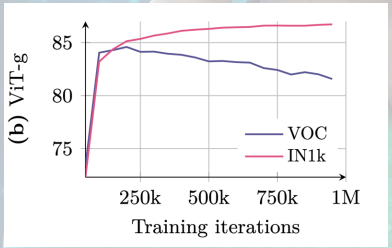
: 성능, 유연성 ↑

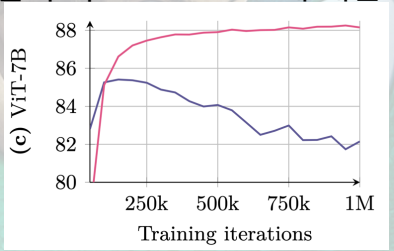
2 Gram Anchoring: A regularization for dense features

- Loss of patch-level consistency over training
- Gram anchoring objective
- Leveraging higher-resolution featuressx

Gram Anchoring– Loss of Patch-Level Consistency Over Training

- 기존 문제 – training을 거듭할 수록 dense task에 대한 성능이 떨어짐



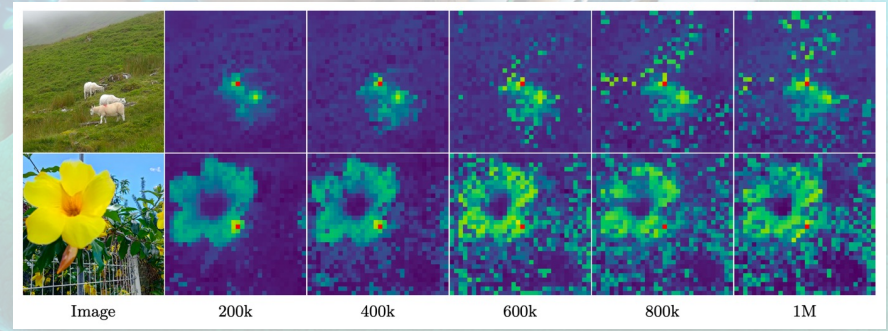


dense
global

segmentation
classification

- patch level data inconsistency 때문

Gram Anchoring– Loss of Patch-Level Consistency Over Training



Gram Anchoring- Gram Anchoring Objective

→ global DINO loss + local iBOT loss를 결합해서 사용했지만, 학습이 반복될 수록 글로벌 표현이 우세

- 새로운 goal

: global performanc는 유지하면서 patch feature들을 regularize하고 patch-level consistency를 보장

Gram Anchoring- Gram Anchoring Objective

- patch-level consistency를 보장할 수 있는 loss 도입
- 이 loss는 Gram Matrix 위에서 동작
- Student의 gram matrix를 Gram teacher의 초기 모델에 집어넣는 것이 목표
- 여기서 early iteration의 teacher model – superior dense properties

Gram Anchoring– Gram Anchoring Objective

step1

$$\mathcal{L}_{\text{Gram}} = \left\| \mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top \right\|_{\text{F}}^2 .$$

$X_s = L2\ normalized\ local\ features\ of\ the\ student$
$X_G = L2\ normalized\ local\ features\ of\ the\ Gram\ teacher$

- Global crop에서만 계산
- 학습의 효율성을 위해 1M iteration 이후에 이 loss를 적용하는데, 늦게 적용하더라도 손실된 dense feature들이 회복 가능

Gram Anchoring– Gram Anchoring Objective

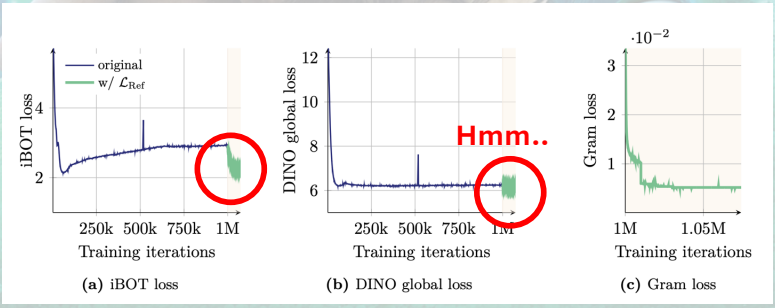
step2

refinement step

$$\mathcal{L}_{\text{Ref}} = w_{\text{D}}\mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{\text{DK}}\mathcal{L}_{\text{DKoleo}} + w_{\text{Gram}}\mathcal{L}_{\text{Gram}}.$$

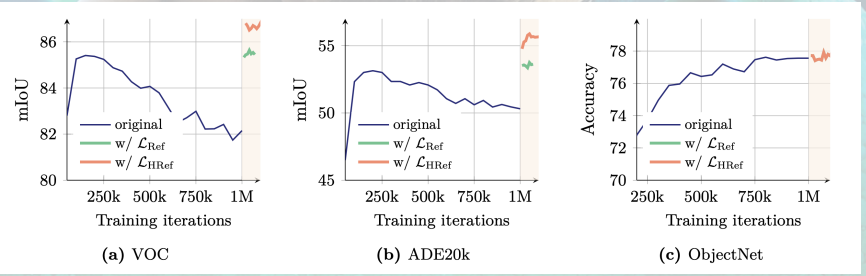
– 성능을 보다 향상시키기 위해, 1M iteration(Gram teacher와 EMA teacher가 동일해지는 타이밍) 마다 Gram teacher update

Gram Anchoring– Gram Anchoring Objective



- Gram Loss와 iBOT Loss는 유사한 방식으로 feature에 영향을 미치지만, Gram Loss와 DINO Loss는 그렇지 않음

Gram Anchoring– Gram Anchoring Objective



성능 ↑ ↑ ↑

Gram Anchoring- Leveraging Higher-Resolution Features

- Gram teacher: Gram anchoring에서 사용되는 초기 단계 model
- 이전에 학습된 model의 Gram matrix를 사용하여 현재 학습 중인 model의 Gram matrix를 교정

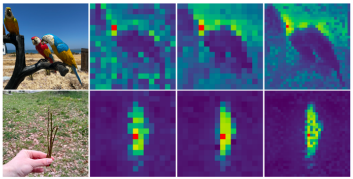
1. feature들의 가중 평균은 outlier patch들을 smoothing하고 consistency를 더 향상시켜, 강력한 local 표현을 생성할 수 있음
2. 고해상도 이미지를 backbone에 입력하면 더 세밀한 feature map이 생성

Gram Anchoring- Leveraging Higher-Resolution Features

- 앞선 1, 2번의 관찰을 기반으로 Gram teacher를 위한 high quality의 feature map을 생성

1. 기존 해상도의 2배인 image를 Gram teacher에게 입력
2. bicubic interpolation으로 2배 downsampling을 하여 student 출력 크기에 맞는 부드러운 feature map을 얻음

Gram Anchoring– Leveraging Higher–Resolution Features



(a) Gram matrices at different input resolutions.

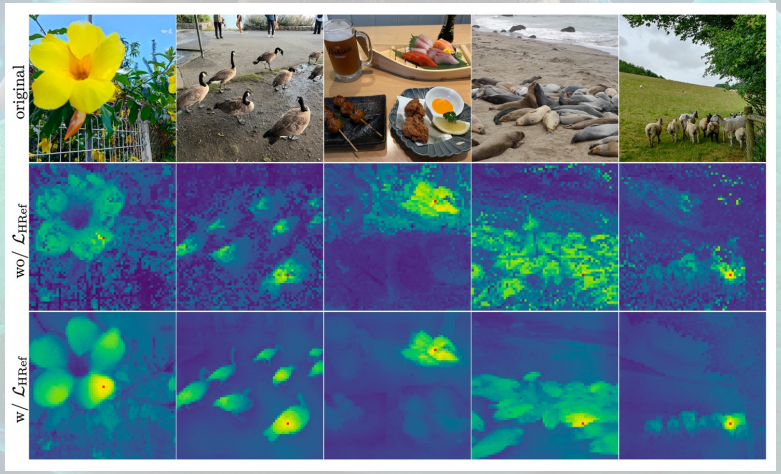
Method	Teacher Iteration	Res.	IN1k Linear	ADE mIoU	NYU RMSE
Baseline	—	—	**88.2**	50.3	0.307
GRAM	200k	×1	88.0	53.6	0.285
GRAM	200k	×2	88.0	**55.7**	**0.281**
GRAM	100k	×2	87.9	**55.7**	0.284
GRAM	1M	×2	88.1	54.9	0.290

(b) Ablation of Gram teachers and resolutions.

Figure 9: Quantitative and qualitative study of the impact of high-resolution Gram. We show (a) the improved cosine maps after down-sampling the high-resolution maps into smaller ones, and (b) the quantitative improvements brought by varying the training iteration and the resolution of the Gram teacher.

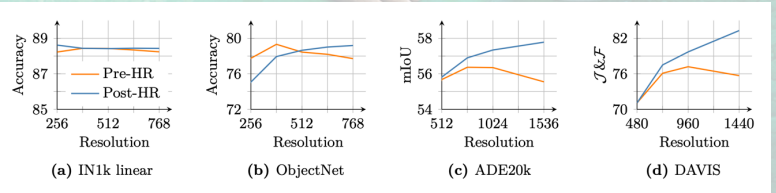
Gram Anchoring- Leveraging Higher-Resolution Features

result

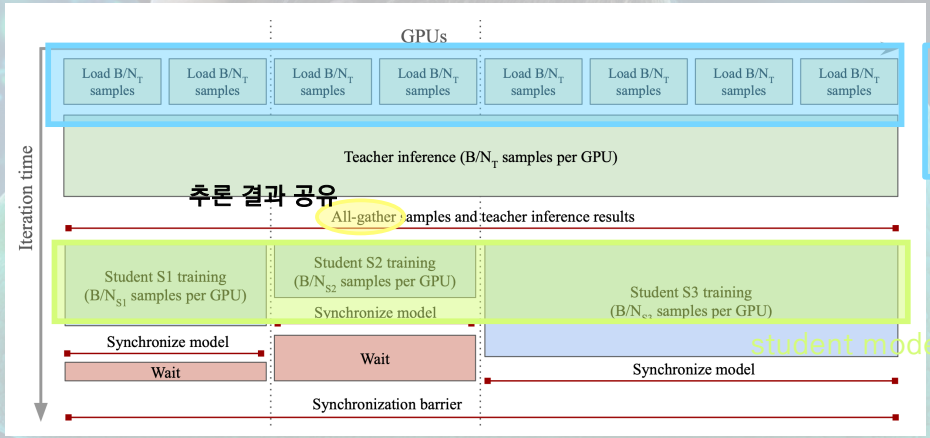


Post Training – Resolution Scaling

- High resolution adaptation 단계의 필요성
  - 지금까지의 model은 256 resolution으로 학습함
  - 하지만 실제 data는 이보다 resolution이 많이 높음
- 7B teacher를 Gram teacher로 사용하는 Gram Anchoring 추가



Post Training – Model Distillation



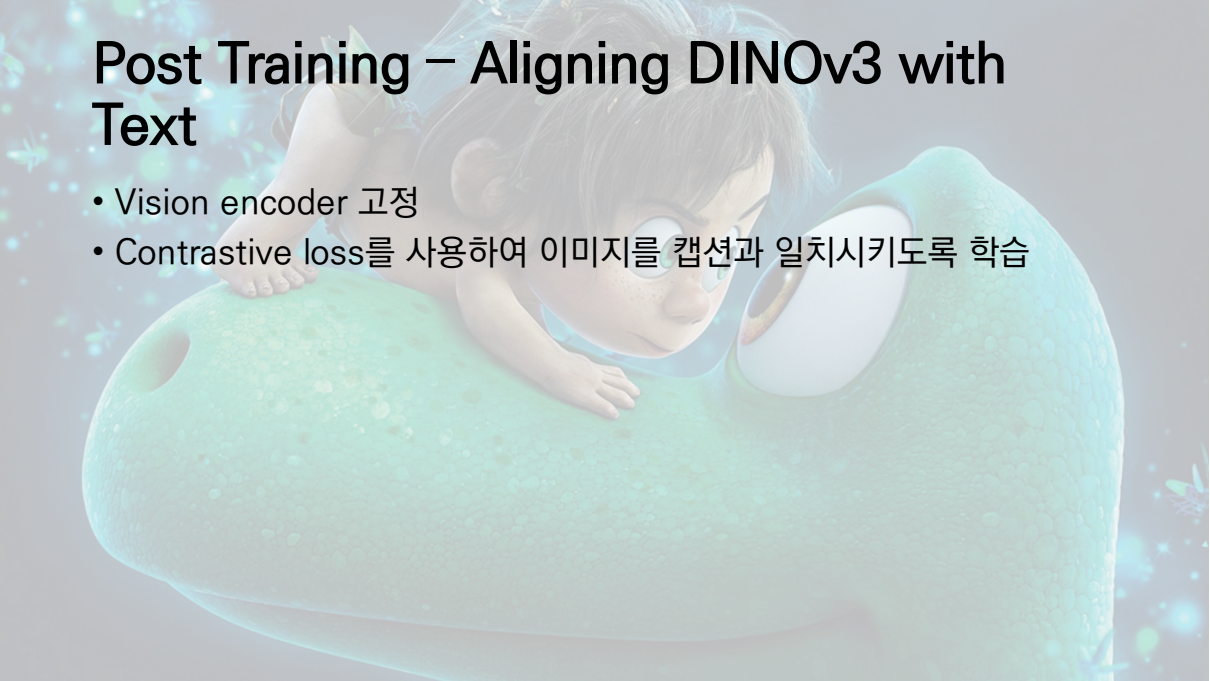
모든 gpu가 teacher model의 추론 실행

student model 훈련

〈multi-student distillation procedure〉

Post Training – Aligning DINOv3 with Text

- Vision encoder 고정
- Contrastive loss를 사용하여 이미지를 캡션과 일치시키도록 학습



Result

- 한 줄 요약 – 성능이 잘 나온다!

