Efficient Self-supervised Vision Transformers for Representation Learning

School of Industrial and Management Engineering, Korea University

Jae Hoon Kim

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ Efficient Self-supervised Vision Transformers for Representation Learning (arXiv, 2021)

- Microsoft에서 연구하였으며 2021년 08월 14일 기준으로 1회 인용됨

Efficient Self-supervised Vision Transformers for Representation Learning

Chunyuan Li[1] **Jianwei Yang**[1] **Pengchuan Zhang**[1] **Mei Gao**[2] **Bin Xiao**[2] **Xiyang Dai**[2]
Lu Yuan[2] **Jianfeng Gao**[1]
[1]Microsoft Research at Redmond, [2]Microsoft Cloud + AI
{chunyl,jianwyan,penzhan,xuga,bixi,xidai,luyuan,jfgao}@microsoft.com

Abstract

This paper investigates two techniques for developing efficient self-supervised vision transformers (EsViT) for visual representation learning. First, we show through a comprehensive empirical study that multi-stage architectures with sparse self-attentions can significantly reduce modeling complexity but with a cost of losing the ability to capture fine-grained correspondences between image regions. Second, we propose a new pre-training task of region matching which allows the model to capture fine-grained region dependencies and as a result significantly improves the quality of the learned vision representations. Our results show that combining the two techniques, EsViT achieves 81.3% top-1 on the ImageNet linear probe evaluation, outperforming prior arts with around an order magnitude of higher throughput. When transferring to downstream linear classification tasks, EsViT outperforms its supervised counterpart on 17 out of 18 datasets. The code and models will be publicly available.

Research Purpose

❖ Efficient Self-supervised Vision Transformers for Representation Learning (arXiv, 2021)

- Vision Transformer 계열 모델로서 self-supervised 방식에 적합한 모델 구조 및 학습 방법론 제안
- 다양한 비교 실험을 통해서 제안한 모델 구조의 타당성을 확립함
- 모델은 patch merging/embedding 모듈과 sparse self-attention 모듈로 구성됨 (Swin transformer 활용)
- 자기지도학습 시 이미지를 view level, region level로 나누어서 학습을 진행
  - ✓ View level은 CLS 토큰을 사용한 loss 계산을 의미, DINO와 동일한 방식으로 진행
  - ✓ Region level은 feature map을 사용한 loss 계산을 의미, dense self-supervised learning 방식으로 진행

** 본 논문을 읽기 전에 **Swin transformer, DINO, dense self-supervised learning**을 읽어보는 것을 추천함

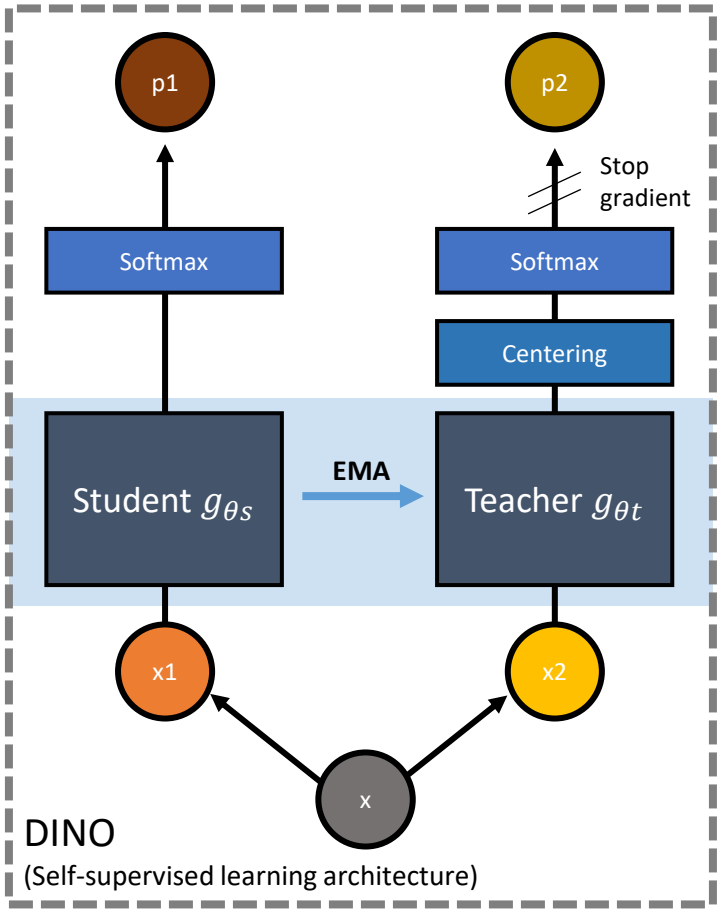
Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *arXiv preprint arXiv:2103.14030* (2021).

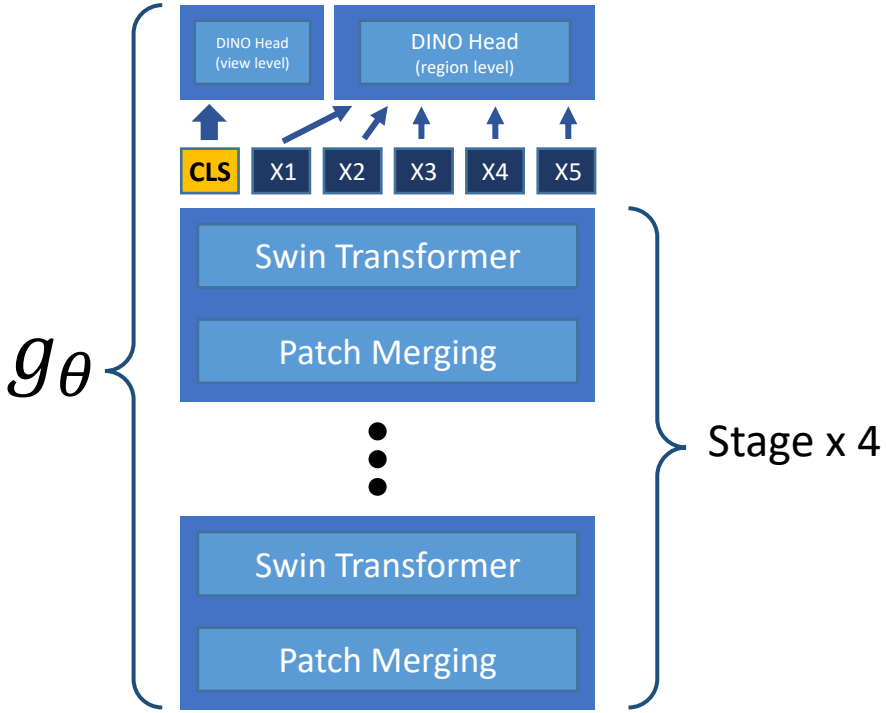
Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *arXiv preprint arXiv:2104.14294* (2021).

Efficient Self-supervised Vision Transformers (EsViT)

❖ Diagram of EsViT

- EsViT의 전체적인 아키텍처는 DINO와 Swin transformer의 조합으로 볼 수 있음





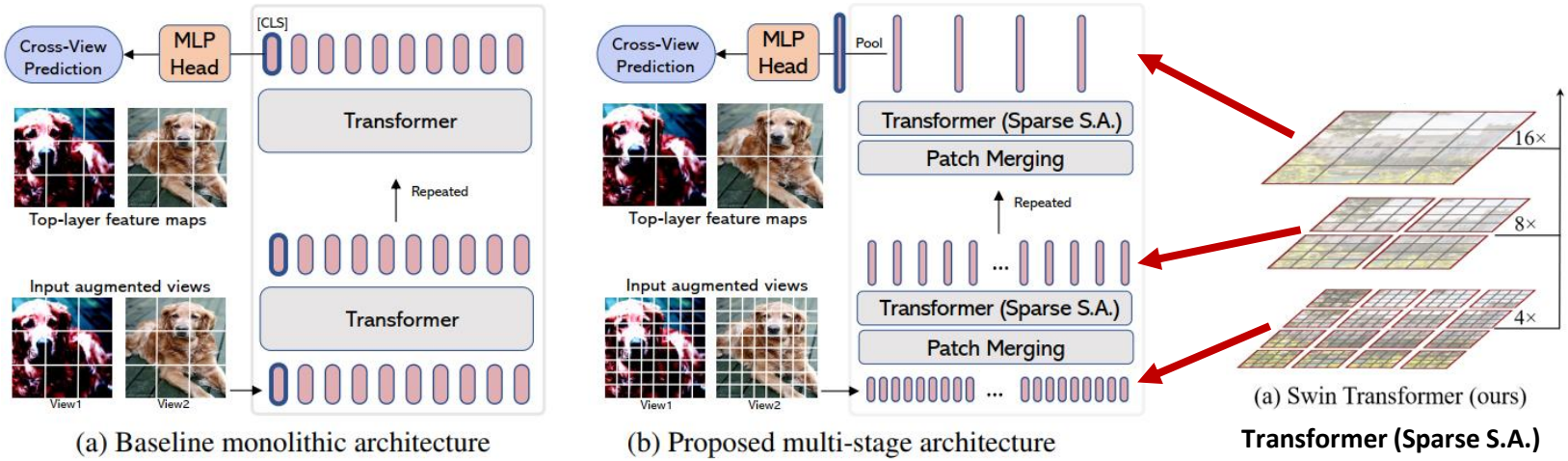
DMQA

Efficient Self-supervised Vision Transformers (EsViT)

❖ Diagram of EsViT (multi-stage ViT)

- EsViT는 back-bone으로 sparse self-attention을 사용하는 transformer 모델을 사용함 (Swin, ViL, CvT)
  - ✓ 비교 실험을 통해서 가장 좋은 성능을 보인 Swin transformer를 최종적으로 사용
- ViT는 훈련이 진행되는 과정 동안 토큰의 크기가 동일하기 때문에 토큰의 수가 바뀌지 않음
- 반면 EsViT는 토큰 간 병합을 통해 토큰의 크기를 점점 키움과 동시에 토큰의 수가 줄어듦
- 따라서 EsViT는 한 토큰이 보는 이미지의 해상도가 점점 작아지는 계층적 구조를 생성함

** Transformer with sparse attention에 대한 내용은 Swin transformer를 참고할 것 → URL [SWIN TRANSFORMER]



< ViT > < EsViT >

DMQA

Efficient Self-supervised Vision Transformers (EsViT)

❖ Diagram of EsViT (pre-training tasks)

- Self-supervised learning에 사용할 데이터셋은 DINO와 동일한 방식으로 생성함
  - ✓ Global token – 원본 이미지 크기의 50% 이상으로 2개 생성 (224 x 224)
  - ✓ Local token – 원본 이미지 크기의 50% 미만으로 8개 생성 (96 x 96)
- Swin Transformer로부터 나온 CLS 토큰은 DINO에서 사용하는 MLP Head의 입력 값으로 사용함
- View-level prediction loss ($\mathcal{L}_V$)는 DINO에서 계산하는 loss와 동일함

  ** DINO에 대한 설명은 다음 URL을 참고할 것 → URL [DINO]
- Region-level prediction loss ($\mathcal{L}_R$)는 multi-stage ViT에서 view-level prediction loss를 계산할 때 발생하는 단점을 보완하기 위해 본 논문에서 제안하는 loss function

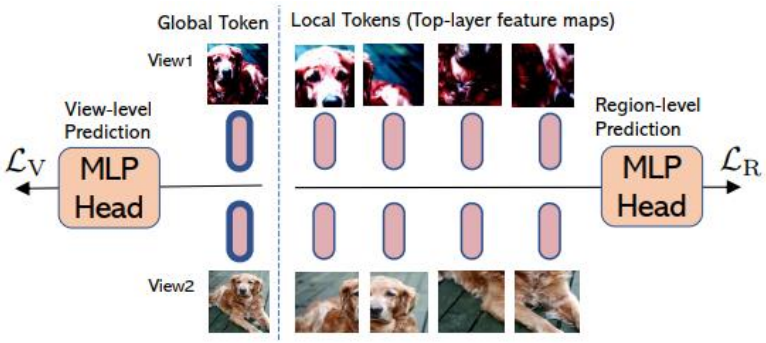


Figure 3: Pre-training objectives.

$$\mathcal{L}_V = \frac{1}{|\mathcal{P}|} \sum_{(s,t)\in\mathcal{P}} -p_s \log p_t$$

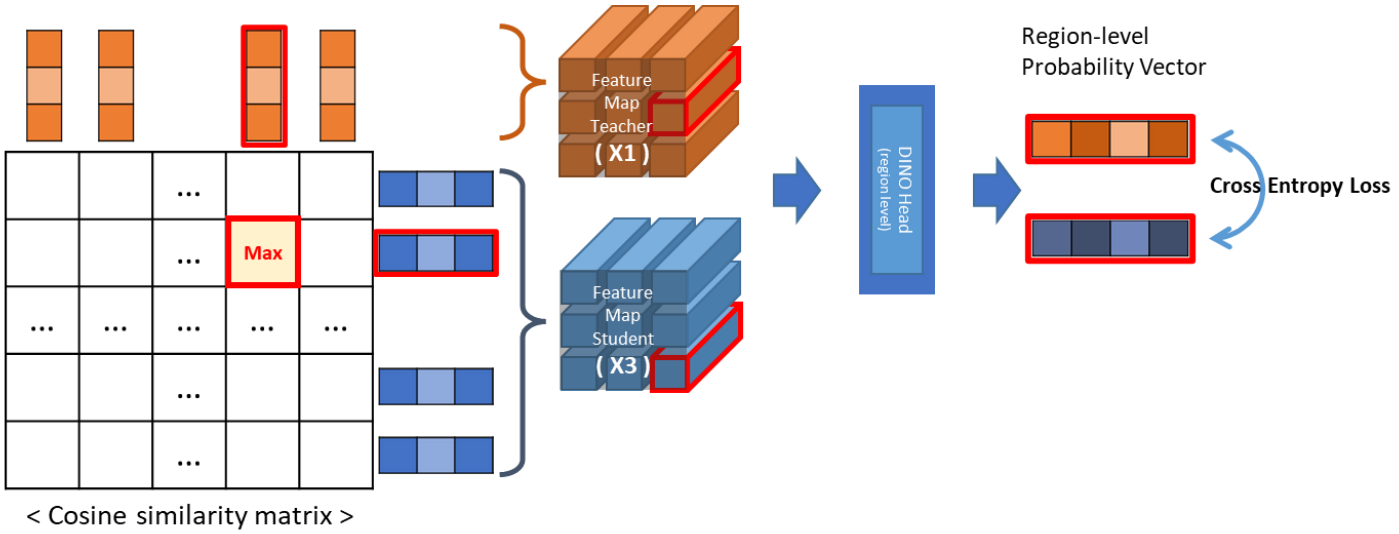
$$\mathcal{L}_R = -\frac{1}{T|\mathcal{P}|} \sum_{(s,t)\in\mathcal{P}} \sum_{i=1}^{T} p_{j^*} \log p_i \qquad j^* = \arg\max_j \frac{z_i^T z_j}{||z_i||\,||z_j||}$$

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_V \text{ (total loss)}$$

Efficient Self-supervised Vision Transformers (EsViT)

❖ Pre-training tasks (region-level task)

- Dense self-supervised 방법론을 활용하여 loss를 계산함
- Teacher와 student networks에서 나온 feature map의 local feature 간 코사인 유사도를 구함
- 그 중 가장 높은 유사도를 보이는 조합의 local feature에 해당하는 probability vector로부터 cross entropy loss를 구함 (region-level prediction loss; $\mathcal{L}_R$)



Experiments

❖ Comparisons with prior art on ImageNet

- 해당 실험에서는 ImageNet 데이터셋으로 기존에 나온 모델들과 성능을 비교함
  - ✓ Transformer based self-supervised learning
  - ✓ Convolution networks based self-supervised learning
  - ✓ Longer sequences for self-attentions
- EsViT에 적합한 back bone 모델을 찾기 위한 비교 실험 진행
- Region level task의 효과를 입증하는 실험 진행

❖ Default training settings (EsViT)

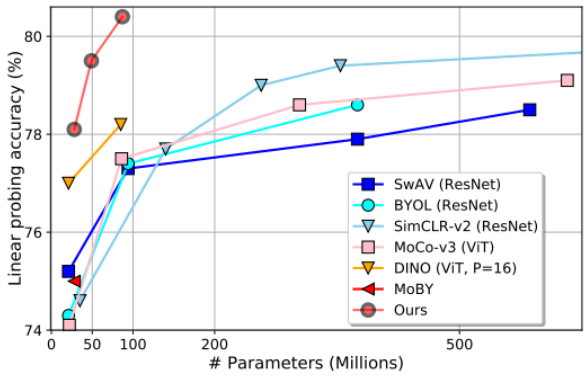
- Dataset: ImageNet-1k
- Optimizer: AdamW
- Batch size: 512
- Learning rate: 0.0005 * (batchsize / 256)
- Learning rate scheduler: Cosine scheduling with linear warm-up (10 epochs)

Experiments

❖ Comparisons with prior art on ImageNet

- Transformer based self-supervised learning
  - ✓ ImageNet validation set으로 linear evaluation protocol 방식으로 학습된 모델의 성능을 비교함
  - ✓ MoBY의 경우, 같은 back bone을 사용하였을 때 EsViT 방식이 더 효과적인 것으로 보임
  - ✓ 전반적인 결과를 볼 때, EsViT가 모델 파라미터 수 대비 더 효과적인 성능을 보이고 있음

Method	#Parameters ↓	Throughput (Image/s) ↑	Linear ↑	k-NN ↑
Transformer-based SSL, with moderate sequence length for self-attentions				
Masked Patch Pred., ViT-B/16 [19]	85	312	79.9†	-
DINO, DeiT-S/16 [6]	21	1007	77.0	74.5
DINO, ViT-B/16 [6]	85	312	78.2	76.1
MoCo-v3, ViT-B/16 [12]	85	312	76.7	-
MoCo-v3, ViT-H-BN/16 [12]	632	~32	79.1	-
MoBY, Swin-T [64]	28	808	75.1	-
EsViT, Swin-T (W=7)	28	808	78.1	75.7
EsViT, Swin-S (W=7)	49	467	79.5	77.7
EsViT, Swin-B (W=7)	87	297	**80.4**	**78.9**



DMQA

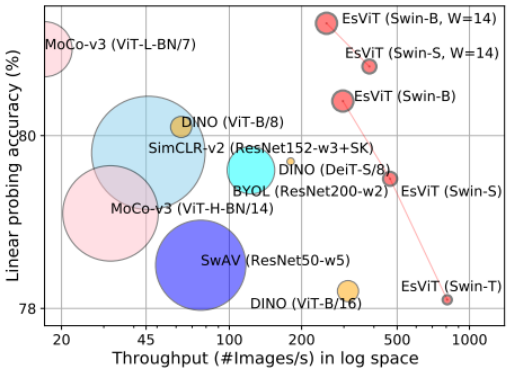
Experiments

❖ Comparisons with prior art on ImageNet

- Longer sequences for self-attentions
  - ✓ EsViT는 Swin transformer를 사용하기 때문에 window 크기를 늘리게 되면 attention 계산량이 늘어남
  - ✓ DINO와 MoCo v3는 ViT를 사용하기 때문에 patch 크기를 작게 할 수록 attention 계산량이 늘어남
  - ✓ 기존 패치/윈도우 크기일 때의 처리량과 비교해볼 때 EsViT가 보다 계산 효율적인 모델임

Method		#Parameters ↓	Throughput (Image/s) ↑	Linear ↑	k-NN ↑
Skyline methods with excessively long sequences for self-attentions					
DINO, DeiT-S/8 [6]	16 → 8	21	180 (1007)	79.7	78.3
DINO, ViT-B/8 [6]		85	63 (312)	80.1	77.4
MoCo-v3, ViT-B-BN/7 [12]	16 → 7	85	∼63 (312)	79.5	-
MoCo-v3, ViT-L-BN/7 [12]		304	∼17	81.0	-
iGPT, iGPT-XL [8]	Patch Size	6801	-	72.0	-
EsViT, Swin-S/$W=14$	7 → 14	49	383 (467)	80.8	79.1
EsViT, Swin-B/$W=14$	Window Size	87	254 (297)	**81.3**	**79.3**

Throughput results
Moderate sequences for self-attentions



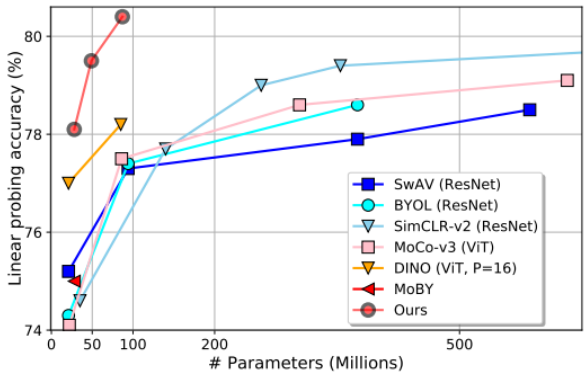
DMQA

Experiments

❖ Comparisons with prior art on ImageNet

- Convolution networks based self-supervised learning
  - ✓ EsViT가 Convolution 기반 모델 보다 훨씬 적은 파라미터 수를 가지면서 보다 높은 성능을 가짐
  - ✓ 또한 데이터 처리(throughput) 속도면에서도 EsViT가 더 우수함

Method	#Parameters ↓	Throughput (Image/s) ↑	Linear ↑	k-NN ↑
SoTA SSL methods with Big ConvNets				
SwAV, RN50w5 [5]	586	76	78.5	67.1
BYOL, RN200w2 [25]	250	123	79.6	73.9
SimCLR-v2, RN152w3+SK [10]	794	46	79.8	73.1
EsViT, Swin-S/$W=14$	49	383	80.8	79.1
EsViT, Swin-B/$W=14$	87	254	**81.3**	**79.3**
EsViT, Swin-T (W=7)	28	808	78.1	75.7
EsViT, Swin-S (W=7)	49	467	79.5	77.7
EsViT, Swin-B (W=7)	87	297	**80.4**	**78.9**

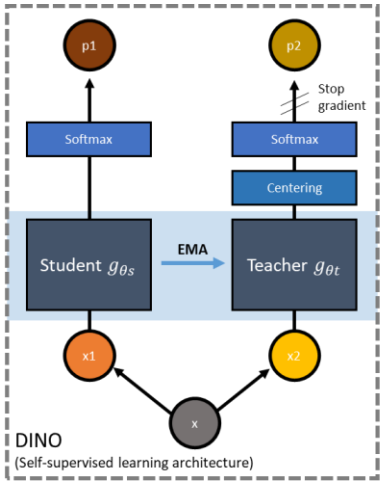


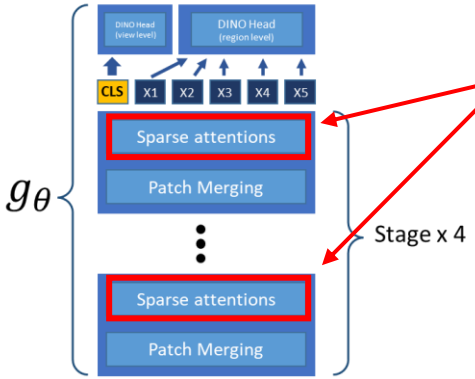
DMQA

Experiments

❖ Comparisons with prior art on ImageNet

- Comparision of sparse attentions
  - ✓ Sparse attention을 사용하는 네 가지 모델에 대하여 성능 비교 실험을 함
  - ✓ 결과적으로 Swin transformer을 사용했을 때 가장 좋은 성능을 보임





Method	#Param.	Im./s	100 epochs Linear	100 epochs k-NN	300 epochs Linear	300 epochs k-NN
DeiT	21	1007	73.1	69.0	75.9	73.2
Swin	28	808	75.3	70.0	77.1	73.7
ViL	28	386	75.4	70.1	77.3	73.9
CvT	29	848	75.5	70.6	77.6	74.8

Table 3: Different sparse attentions in EsViT.

Experiments

❖ Comparisons with prior art on ImageNet

- The effectiveness of adding region-level task
  - ✓ View-level task로만 학습을 진행했을 경우보다 region-level task와 함께 진행했을 때의 성능이 더 높음
  - ✓ View-level task는 입력 이미지의 전반적인 특징 정보를 가진 CLS 토큰으로 학습이 진행되는데,
  - ✓ 이미지의 global invariance feature만으로도 충분하기 때문에 local invariance feature를 따로 추출하지 않음
  - ✓ 다만, DINO에서는 별도의 region-level task가 없더라도 학습 성능이 떨어지지 않기 때문에 multi-stage architecture가 문제일 것이라고 논문에서는 추측함

* 여기서 이야기하는 global, local은 입력된 이미지의 전체와 일부분을 의미하는 것으로 추정 (DINO, SwAV에서의 global, local 아님)

Arch.	Objectives	Window S.	Linear	k-NN
Swin-T	$\mathcal{L}_V$	7	77.0	74.2
	$\mathcal{L}_V+\mathcal{L}_R$	7	**78.1**	**75.7**
	$\mathcal{L}_V$	14	77.9	75.5
	$\mathcal{L}_V+\mathcal{L}_R$	14	**78.7**	**77.0**
Swin-S	$\mathcal{L}_V$	7	79.2	76.8
	$\mathcal{L}_V+\mathcal{L}_R$	7	**79.5**	**77.7**
	$\mathcal{L}_V$	14	79.4	77.3
	$\mathcal{L}_V+\mathcal{L}_R$	14	**80.8**	**79.1**
Swin-B	$\mathcal{L}_V$	7	79.6	77.7
	$\mathcal{L}_V+\mathcal{L}_R$	7	**80.4**	**78.9**
	$\mathcal{L}_V$	14	80.5	78.3
	$\mathcal{L}_V+\mathcal{L}_R$	14	**81.3**	**79.3**

Table 2: Ablations of pre-train tasks and window sizes.

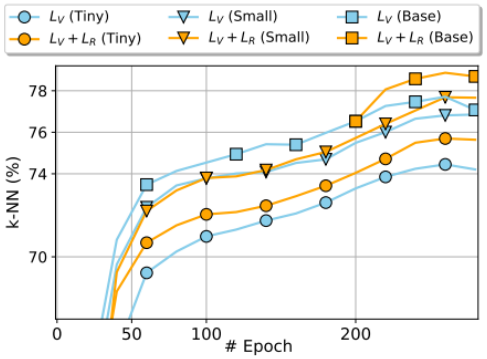


Figure 4: Learning curve of different pre-training tasks. For Base model, $\mathcal{L}_R$ is added from the 200th epoch.

Experiments

❖ Qualitative studies

- Visualization of correspondences
  - Region-level task($\mathcal{L}_R$)를 수행할 경우 local feature에 대한 정보를 학습하기 때문에 EsViT가 local matching 문제에서 좋은 성능을 보일 수 있도록 함
  - View-level task($\mathcal{L}_V$)만 학습한 경우 local matching 문제에서 성능이 떨어지는 것을 확인할 수 있음

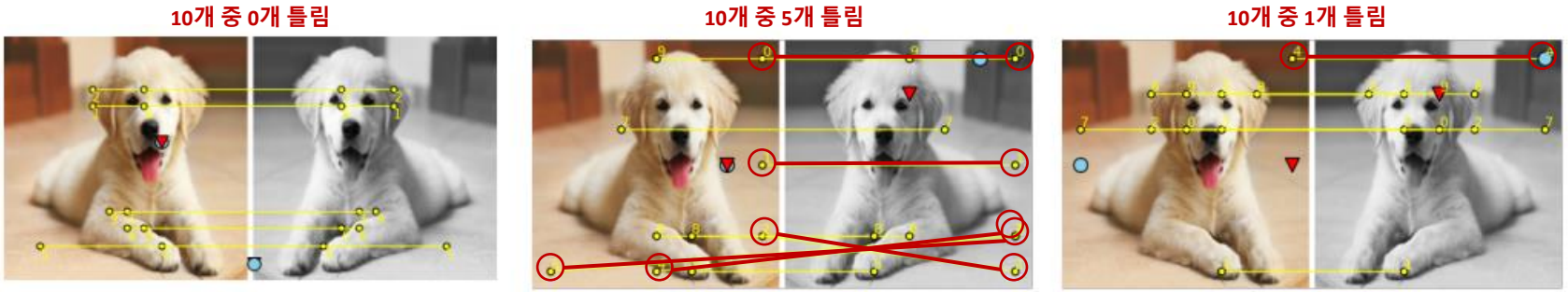
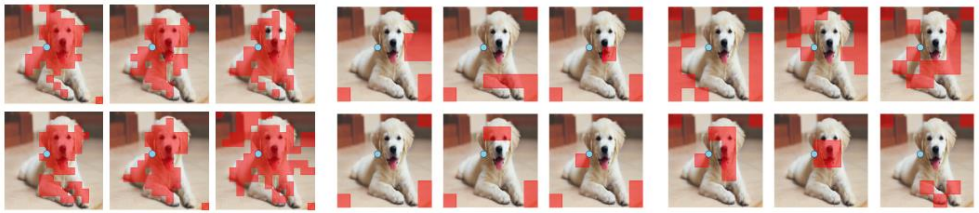


Figure 6: The learned correspondences. Yellow lines are the top-10 correspondences between two views, where the numbers indicates the rankings of similarity scores, yellow dots with the same number are paired. The blue dot and red triangle indicates the most similar local regions that correspond to the global feature of the view itself and the other view, respectively.

Experiments

❖ Qualitative studies

- Visualization of attention maps
  - DINO와 같이 multi-stage architectur가 아닌 경우에는
  - 이는 고양이 그림에서 query를 배경으로 주었음에도 attention이 객체에 집중되는 것으로 알 수 있음
  - 한편, 강아지 그림에서 확인할 수 있듯이 multi-stage architecture를 view-level task로만 진행할 경우 이미지의 correspondence를 자동으로 배우는 특성을 잃어버리게 됨 (이 때 query는 강아지를 가리킴)
  - 따라서 region-level task를 함께 진행하면서 local feature를 보다 잘 학습하여 correspondence 인지 성능을 높이는 것이 중요함



(a) DINO: DeiT-S (b) EsViT: $\mathcal{L}_V$ (c) EsViT: $\mathcal{L}_V + \mathcal{L}_R$

Figure 7: Visualization of the the learned attention map for different heads in the last layer. The query is the **blue** dot in the center of the images. We visualize masks (as **red**) obtained by thresholding the self-attention maps to keep 60% of the probability mass. Note that all 6 heads are visualized for DINO with DeiT-S, and 6 out of 24 heads in EsViT are chosen to visualize (ranked by entropy values). Please see enlarged pictures with all heads in Appendix.

(a) DINO: DeiT-S (b) EsViT: $\mathcal{L}_V$ (c) EsViT: $\mathcal{L}_V + \mathcal{L}_R$

Figure 12: The learned attention maps for all heads at the top layer, ranked by the entropy of softmax probability. Query is the blue dot in the top-left of the image. Top: Entropy of each heads. Middle: top 60% probability mass. Bottom: full attention maps. DINO mainly attends the main object even when the query is a background region.

Conclusion

- ViT 관련 최신 모델의 조합으로 효율적인 연산과 높은 성능을 동시에 달성한 모델 개발
- 이론적인 접근보다 다양한 실험을 통해서 모델 구조를 검증함

Reference

1. Li, Chunyuan, et al. "Efficient Self-supervised Vision Transformers for Representation Learning." *arXiv preprint arXiv:2106.09785* (2021).
2. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *arXiv preprint arXiv:2103.14030* (2021).
3. Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *arXiv preprint arXiv:2104.14294* (2021).
4. Wang, Xinlong, et al. "Dense contrastive learning for self-supervised visual pre-training." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.


Thank You