SiT:
Self-supervised vIsion Transformer

School of Industrial and Management Engineering, Korea University

Sae Rin Lim

KOREA UNIVERSITY

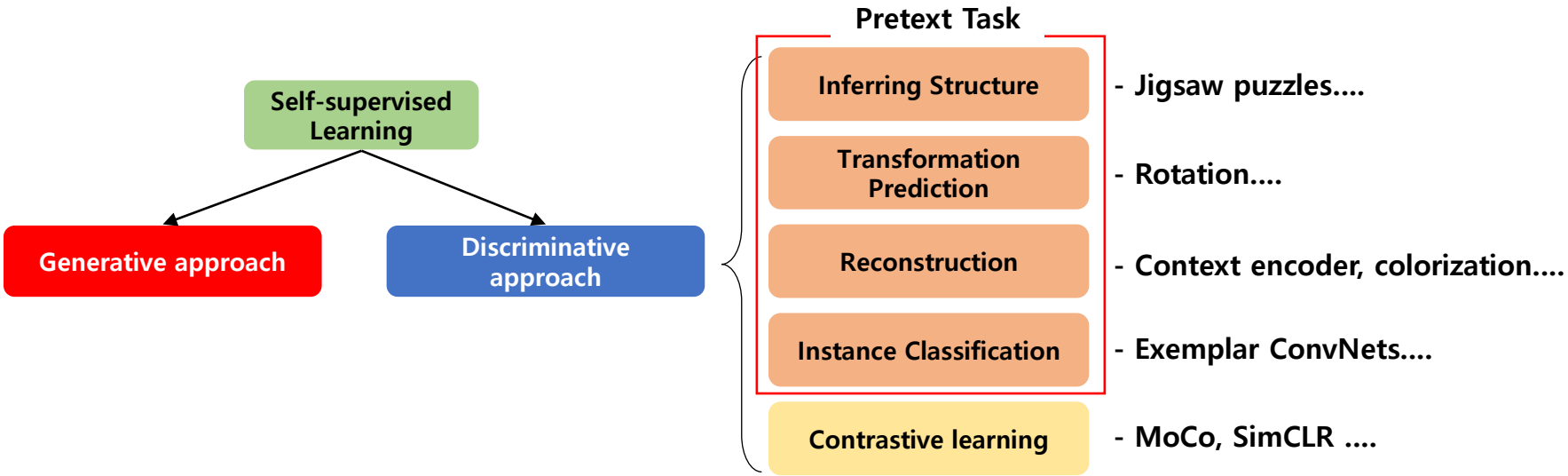
DMQA

• Contents

• Background : Self-supervised Learning

❖ Self-supervised Learning

- labeled data 수집의 어려움과 수많은 unlabeled data를 활용하기 위한 비지도 학습 방법론 중 하나의 분야
- 스스로(self) label을 생성하여(supervision) 사전학습을 하고 이를 소량의 label data를 활용해 downstream task로 전이학습을 하는 접근 방법
- Vision분야에서는 **Self-supervised Learning**을 이용하여 **이미지에 대한 general한 representation**을 얻는 것이 목표
- **Self-supervised learning**은 크게 **generative approach**와 **discriminative approach**로 나눌 수 있으며 discriminative approach는 다시 **Pretext task**와 **Contrastive learning**으로 분류할 수 있음

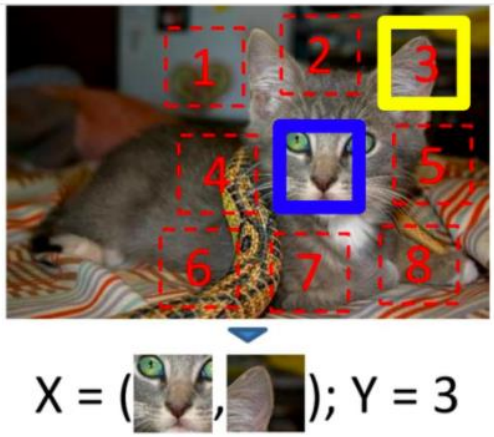


DMQA

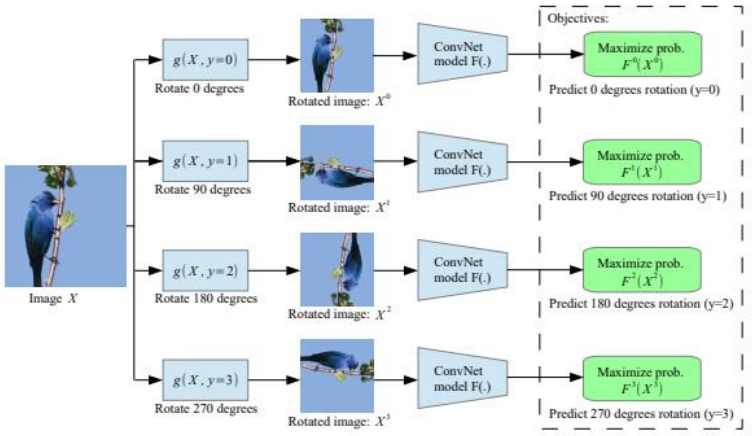
• Background : Self-supervised Learning

❖ Pretext Task

- GAN을 기반으로 한 Generative approach는 학습 불안정성과 data distribution 학습에 대한 계산 자원 등으로 비효율적이며 representation learning에 필수적이지 않을 수도 있음(저자들의 주장)
- 때문에 Discriminative approach가 주류가 되었으며 초기에는 **pretext task**를 활용한 방법론이 주를 이룸
- Pretext task란 사용자가 이미지(입력변수)를 활용한 새로운 문제를 정의하고 모델이 그 문제를 푸는 방향으로 학습을 하는 방법론
- Pretext Task는 문제를 해결하는 과정에서 이미지에 대한 general representation, 저자들의 표현을 빌리면 visual integrity(시각적 온전함)을 학습할 것이라는 가정을 가지고 있음
  - Ex. Rotation 에서 가정 : 모델이 이미지에서 **배경과 객체에 대한 개념**을 알아야 그 객체에 대한 방향을 예측할 수 있을 것이다



Context prediction



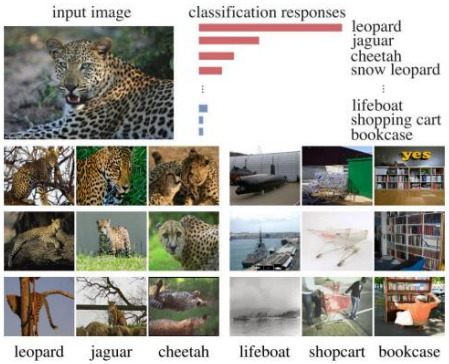
Rotation prediction

DMQA

• Background : Self-supervised Learning

❖ Contrastive Learning

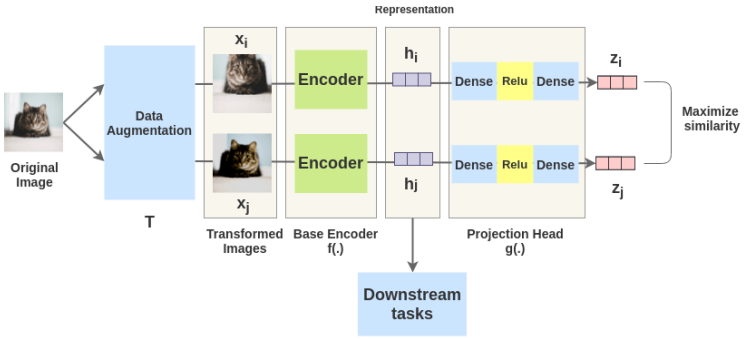
- Contrastive Learning은 Classification Task에서 비슷한 이미지가 높은 확률을 가지는 현상을 확인하고 **비슷한 이미지에 유사한 representation이 있을 것**이라는 가정으로 시작
- **Instance classification**을 Pretext Task로 구성해 augmentation한 이미지를 같은 class로 분류하는 방법론이 존재했지만 각각의 이미지가 하나의 클래스가 되어버려(이미지의 수 = 클래스의 수) 대용량 데이터에 부적합한 문제 발생
- 이를 해결하기 위해 **multi classification 문제를 binary classification 문제로 치환**하여 푸는 **Contrastive Learning**이 발전
- **Contrastive Learning**은 Positive pair와 Negative pair를 정의하여 수많은 Negative pair중 하나의 Positive pair를 찾는 binary classification Task로 많은 논문에서 그 효과를 입증하며 SOTA를 달성



[지도학습 기반 이미지 분류 모델 결과]

Motivation of Contrastive Learning
[출처 : http://dmqm.korea.ac.kr/activity/seminar/302]

Instance Classification : Exemplar



Contrastive Learning : SimCLR
[출처 : https://amitness.com/2020/03/illustrated-simclr/]

DMQA

• Overview of the SiT

❖ SiT : Self-supervised Vision Transformer

- Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, United Kingdom 에서 연구
- 2021년 08월 04일 기준 2회 인용
- **Vision Transformer를 활용한 self-supervised learning 방법론 개발**

SiT: Self-supervised vIsion Transformer

Sara Atito, *Member IEEE,* Muhammad Awais, and Josef Kittler, *Life Member, IEEE*

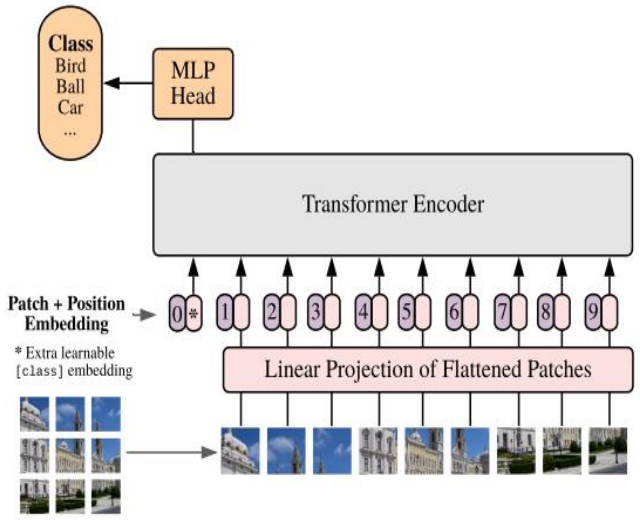
Abstract—
Self-supervised learning methods are gaining increasing traction in computer vision due to their recent success in reducing the gap with supervised learning. In natural language processing (NLP) self-supervised learning and transformers are already the methods of choice. The recent literature suggests that the transformers are becoming increasingly popular also in computer vision. So far, the vision transformers have been shown to work well when pretrained either using a large scale supervised data [1] or with some kind of co-supervision, e.g. in terms of teacher network. These supervised pretrained vision transformers achieve very good results in downstream tasks with minimal changes [1], [2], [3]. In this work we investigate the merits of **self-supervised learning** for pretraining image/vision transformers and then using them for downstream classification tasks. We propose **S**elf-supervised **vi**sion **T**ransformers (**SiT**) and discuss several self-supervised training mechanisms to obtain a pretext model. The architectural flexibility of SiT allows us to use it as an autoencoder and work with multiple self-supervised tasks seamlessly. We show that a pretrained SiT can be finetuned for a downstream classification task on small scale datasets, consisting of a few thousand images rather than several millions. The proposed approach is evaluated on standard datasets using common protocols. The results demonstrate the strength of the transformers and their suitability for self-supervised learning. We outperformed existing self-supervised learning methods by large margin. We also observed that SiT is good for few shot learning and also showed that it is learning useful representation by simply training a linear classifier on top of the learned features from SiT. Pretraining, finetuning, and evaluation codes will be available under: https://github.com/Sara-Ahmed/SiT.

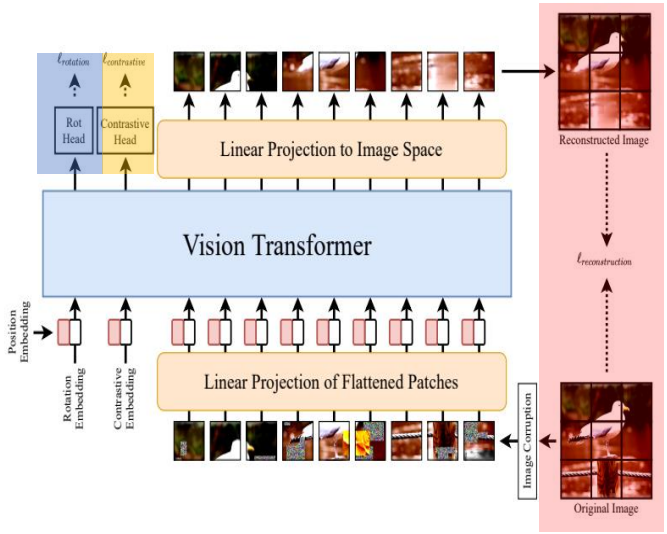
Index Terms—Vision Transformer, Self-supervised Learning, Discriminative Learning, Image Classification, transformer based autoencoders.

• Key points of the SiT

❖ Key points

- **Vision Transformer 구조의 장점을 이용해 multi self-supervised task로 더욱 general한 representation 학습**
  - ➢ Transformer encoder output에 Token을 추가하거나 단순히 FC layer를 쌓는 것 만으로도 다양한 Task를 할 수 있다는 장점을 이용하여 **여러 self-supervised discriminative approach**를 수행
  - ➢ Vision Transformer의 Classification Token을 제거하고 Rotation Token과 Contrastive Token을 추가하여 **Rotation task**와 **Contrastive learning**을 진행
  - ➢ Vision Transformer의 output에 FC layer를 쌓아 CNN-based Autoencoder보다 효율적인 Autoencoder효과를 가지는 모델 구조를 이용하여 **Reconstruction task**에 이용



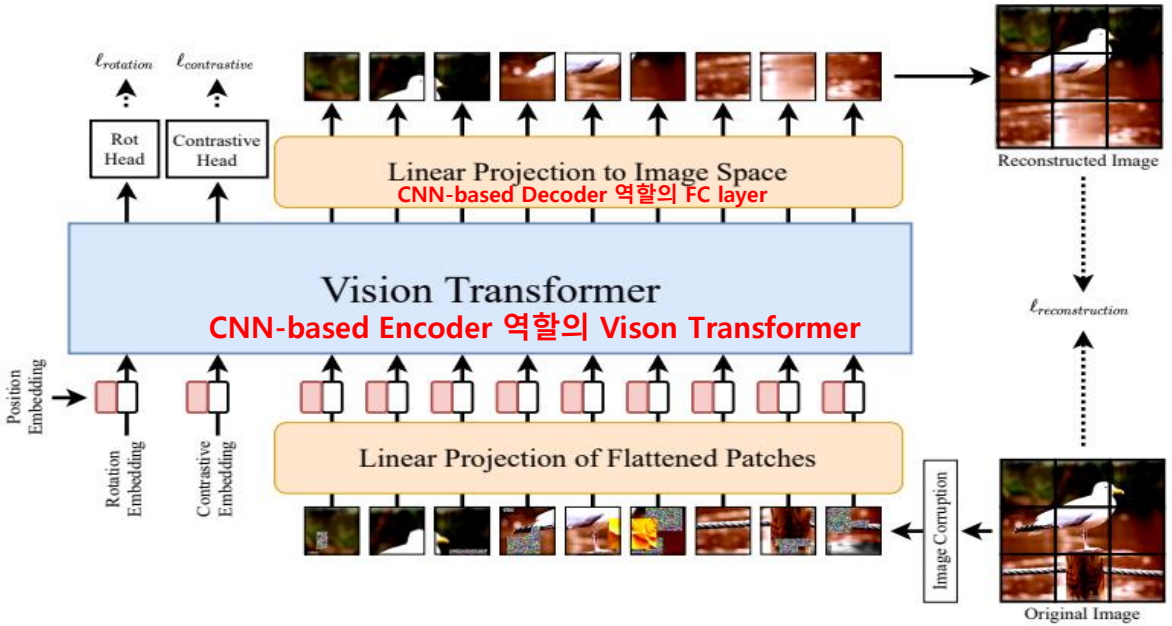


DMQA

• Key points of the SiT

❖ Transformer as autoencoder

- 기존 CNN-based autoencoder는 Conv layer 또는 TransConv layer로 구성된 계산 자원이 높은 Decoder가 필요함
- 또한 Encoder에는 pooling layer나 convolution kernel의 stride에 의한 정보요약 과정에서 버려지는 정보가 발생함
- Transformer encoder의 마지막 **encoder block에 단순히 FC layer를 쌓는 것으로 Decoder 대체**할 수 있으며 Encoder에 convolution이나 pooling층이 없어 정보의 손실 없이 Autoencoder처럼 작동하는 모델 구조 제안

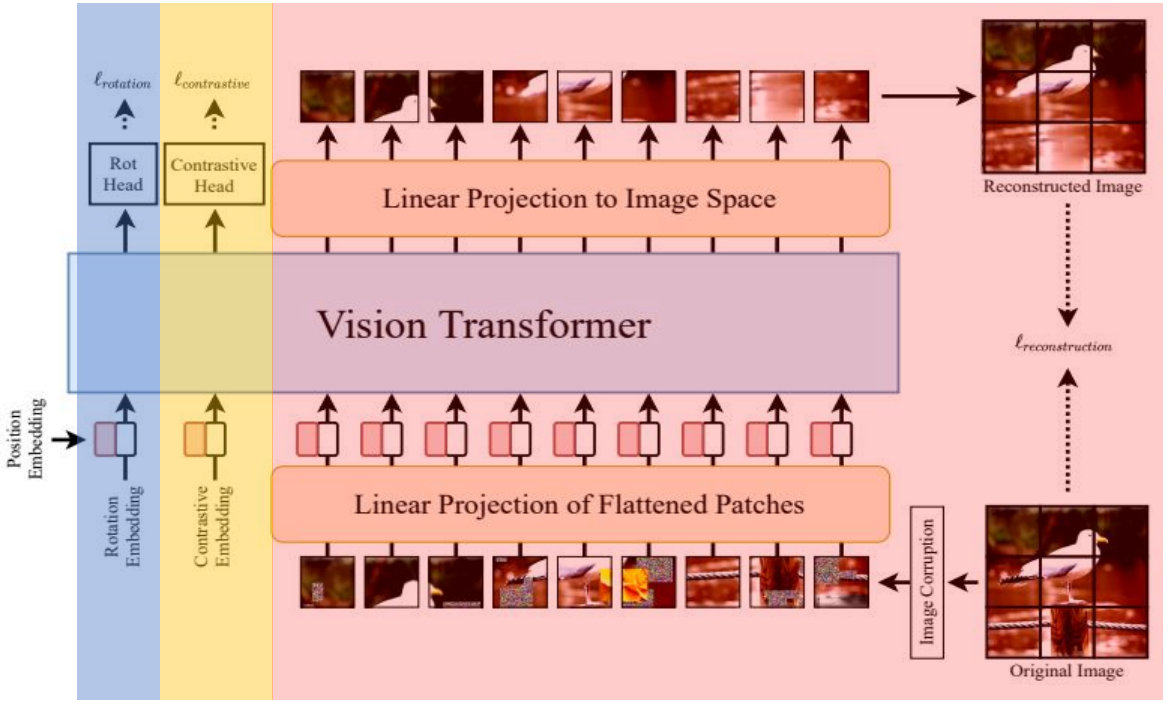


DMQA

• Key points of the SiT

❖ Multi Self-supervised Task

- **Transformer encoder구조의 장점을 이용해 multi self-supervised task로 더욱 general한 representation 학습**
  - ➢ 하나의 task를 통해 얻을 수 있는 representation은 해당 task의 변형과 관련된 것만 얻을 수 있다는 한계 존재
  - ➢ 여러 논문에서 multi self-supervised task를 통해 성능향상을 가지고 올 수 있음을 확인
  - ➢ **Pretext Task(Rotation + Reconstruction)와 Contrastive Learning (SimCLR Framework)을 동시에 활용하는 방법론 제안**



DMQA

• Key points of the SiT

❖ Pretext Task : Reconstruction & Rotation

- **Reconstruction**
  - Vision Transformer를 autoencoder 처럼 활용하여 input image와 reconstructed image에 대한 차이를 줄이며 학습
  - 더 좋은 representation을 얻기 위해서 patch에 다양한 **local transformation**을 적용(random drop, random replace 등)
  - BERT에서 하나의 Token에 Mask를 씌우는 것과 다르게 주변 patch와 연결된 local transformation을 사용
  - Reconstruction Loss는 아래와 같이 pixel단위 $\ell 1$-norm 을 이용하여 계산
  - $\mathcal{L}_{\text{recons}}(\mathbf{W}) = \frac{1}{N}\sum_{i}^{N} ||\mathbf{x_i} - \text{SiT}_{\text{recons}}(\bar{\mathbf{x}}_\mathbf{i})||$
- **Rotation**
  - Vision Transformer의 Classification Token을 제거한 뒤, 새로 추가한 Rotation Token을 통해 예측값 출력
  - Rotation Loss는 아래와 같이 Cross entropy로 계산
  - $\mathcal{L}_{\text{rotation}}(\mathbf{W}) = -\frac{1}{N}\sum_{i}^{N} \log \hat{y}^{\theta}$

Local Transformation 예시

• Key points of the SiT

❖ Contrastive Learning & Final Loss Function

- **Contrastive Learning**
  - ➢ SimCLR를 기반으로, 한 이미지에 서로 다른 augmentation을 적용한 뒤, cosine similarity를 통해 contrastive loss 계산
  - ➢ Rotation과 같이 Contrastive Token을 추가하여 이미지에 대한 representation을 출력
  - ➢ SimCLR과 마찬가지로 NT-Xent loss(Normalized Temperature-scaled Cross Entropy Loss)를 통해 아래와 같이 계산
  - ➢ $$\ell_{\text{contr}}^{x_i,x_j}(\mathbf{W}) = \frac{e^{\text{sim}(\text{SiT}_{\text{contr}}(x_i),\ \text{SiT}_{\text{contr}}(x_j))/\tau}}{\sum_{k=1,k\neq i}^{2N} e^{\text{sim}(\text{SiT}_{\text{contr}}(x_i),\ \text{SiT}_{\text{contr}}(x_k))/\tau}}$$
- **Final Loss Function**
  - ➢ 최종적인 Loss function은 세 가지 loss에 가중합을 하여 계산
  - ➢ Loss앞의 가중치를 grid search로 찾지 않고 **uncertainty weighting approach**를 적용하여 아래 식과 같이 학습 가능한 파라미터로 변경

> * Uncertainty weighting approach는 multi-task training을 할 때, scale이 다른 각각의 loss를 단순히 가중합만 하면 가중치에 너무 민감해지는 현상을 막고 가중치를 찾기 위한 grid search에 요구되는 시간적,계산적 자원을 줄이기 위해 연구된 방법론[6]

  - ➢ $$\begin{aligned}\mathcal{L}_{\text{total}}(\mathbf{W},\alpha_1,\alpha_2,\alpha_3) = &\frac{1}{\alpha_1}\times\mathcal{L}_{\text{recons}}(\mathbf{W}) \\ &+\frac{1}{\alpha_2^2}\times\mathcal{L}_{\text{rotation}}(\mathbf{W}) \\ &+\frac{1}{\alpha_3^2}\times\mathcal{L}_{\text{contr}}(\mathbf{W}) \\ &+\log(\alpha_1)+\log(\alpha_2)+\log(\alpha_3)\end{aligned}$$

• Experiments

❖ Experiments : Pre-trained representation

- CIFAR-10, CIFAR-100, Tiny-ImageNet, STL-10 Dataset을 학습에 사용
- Freeze된 Network에 Linear Classifier를 붙여 성능을 검증하는 Linear Evaluation protocol
- 사전학습을 통해 생성된 representation이 linear evaluation에서 좋은 결과를 얻을 것 뿐만 아니라 Domain Transfer에서도 좋은 성능을 얻음

Method	Backbone	Linear Evaluation			Domain Transfer	
		CIFAR10	CIFAR100	Tiny-ImageNet	C100→C10	C10 →C100
DeepCluster [18]	ResNet-32	43.31% ± 0.62	20.44% ± 0.80	11.64%0.21±	43.39% ± 1.84	18.37% ± 0.41
RotationNet [22]	ResNet-32	62.00% ± 0.79	29.02% ± 0.18	14.73%0.48±	52.22% ± 0.70	27.02% ± 0.20
Deep InfoMax [19]	ResNet-32	47.13% ± 0.45	24.07% ± 0.05	17.51%0.15±	45.05% ± 0.24	23.73% ± 0.04
SimCLR [8]	ResNet-32	77.02% ± 0.64	42.13% ± 0.35	25.79%0.4±	65.59% ± 0.76	36.21% ± 0.16
Relational Reasoning [20]	ResNet-32	74.99% ± 0.07	46.17% ± 0.16	30.54%0.42±	67.81% ± 0.42	41.50% ± 0.35
[20]	ResNet-56	77.51% ± 0.00	47.90% ± 0.27	n/a	68.66% ± 0.21	42.19% ± 0.28
SiT (ours)	Transformer	**81.20%**	**55.97%**	**40.67%**	**73.79%**	**55.72%**

TABLE 1: Linear evaluation after self-supervised pretraining. Mean accuracy (percentage) and standard deviation over three runs are reported on CIFAR-10 (C10) and CIFAR-100 (C100) datasets. $X \rightarrow Y$ implies that the pretraining is performed on unlabelled dataset X and the linear evaluation is performed on labeled dataset Y. The best results are highlighted in bold.

• Experiments

❖ Experiments : Few-shot learning

- CIFAR-10, CIFAR-100 Dataset을 통해 few-shot learning에 대한 성능 평가
- 사전 학습된 모델을 사용가능한 labeled data에 finetuning
- 2번째 row는 Few-shot manner로 평가한 지표이며 3번째 row는 linear evaluation protocol로 평가한 지표

Method	0%	1%	10%	25%	50%	100%
CIFAR-10						
[20]	74.99% ± 0.07	76.55% ± 0.27	80.14% ± 0.35	85.30% ± 0.28	89.35% ± 0.11	90.66% ± 0.23
SiT (ours) - fewshot	n/a	74.78%	87.16%	92.90%	94.84%	97.70%
SiT (ours)	81.20%	81.72%	87.90%	93.12%	95.14%	97.53%
CIFAR-100						
[20]	46.17% ± 0.17	46.10% ± 0.29	49.55% ± 0.36	54.44% ± 0.58	58.52% ± 0.70	58.96% ± 0.28
SiT (ours) - fewshot	n/a	27.50%	53.72%	67.58%	74.46%	80.30%
SiT (ours)	55.97%	56.81%	61.35%	70.58%	75.97%	80.20%

TABLE 2: Test accuracy on CIFAR10 and CIFAR100 datasets with respect to the available percentage of the labeled data. SiT is finetuned with the available labeled data (fewshot), followed by linear evaluation on the entire labeled dataset.

• Experiments

❖ Experiments : Finetuning

- STL-10 Dataset에 대해 finetuning한 성능 평가

Method	Backbone	Accuracy
Exemplars [47]	Conv-3	72.80%
Artifacts [23]	Custom	80.10%
ADC [48]	ResNet-34	56.70%
Invariant Info Clustering [49]	ResNet-34	88.80%
DeepCluster [18]	ResNet-34	73.37%
RotationNet [22]	ResNet-34	83.22%
Deep InfoMax [19]	AlexNet	77.00%
Deep InfoMax [19]	ResNet-34	76.03%
SimCLR [8]	ResNet-34	89.31%
Relational Reasoning [20]	ResNet-34	89.67%
SiT (our)	Transformer	**93.02%**

TABLE 3: A comparison with the state-of-the-art methods based on an unsupervised training and finetuning experiment involving the STL-10 dataset

• Experiments

❖ Experiments : Ablation Study

- STL-10 Dataset 사용, 각각의 self-supervised task와 multi self-supervised task의 효과를 알아보기 위한 비교 실험 진행

Self-supervision tasks	**Finetuned**	**Linear Evaluation**
Trained from scratch	59.38%	37.75%
Reconstruction	90.03%	45.49%
Rotation	65.48%	46.80%
Contrastive	84.50%	69.90%
Reconstruction+Rotation	91.80%	70.38%
Reconstruction+Contrastive	89.46%	73.90%
Rotation+Contrastive	90.44%	77.10%
Reconstruction+Rotation+Contrastive	91.49%	**78.58%**
Reconstruction+Rotation+Contrastive (uncertainty weighting)	**93.02%**	78.51%

TABLE 4: Performance of the individual elements of the pretext learning.

• Conclusion

- 각 token이 한 path로만 흐른다는 것과 Output을 자유롭게 구성할 수 있다는 Transformer 구조의 장점을 이용한 multi self-supervised learning 방법론을 제안
- 각각의 task에 대한 ablation study를 통해 그 효과를 입증하고 여러 데이터셋에서 SOTA 성능 달성

- 후기
  - ➢ CNN-based autoencoder를 단순히 FC layer를 쌓아 decoder를 대체한 구조는 Vision transformer를 가장 잘 이용한 pretext task 구성이라고 생각함. 특히 reconstruction 만을 이용한 finetuning 성능이 contrastive learning 성능보다 뛰어난 것이 인상적임
  - ➢ 연구실 단위에서 진행된 실험이라 ImageNet과 같이 큰 데이터셋에 대한 실험이 없어 Context Encoder와 같은 방법론과 직접적으로 비교하지 못한 것이 아쉬움

• Reference

1. Atito, S., Awais, M., & Kittler, J. (2021). Sit: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
3. Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
4. Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision* (pp. 1422-1430).
5. Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), 1734-1747.
6. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7482-7491).

KOREA UNIVERSITY

DMQA