최신 자가 학습 기반의 인공지능 기술 동향

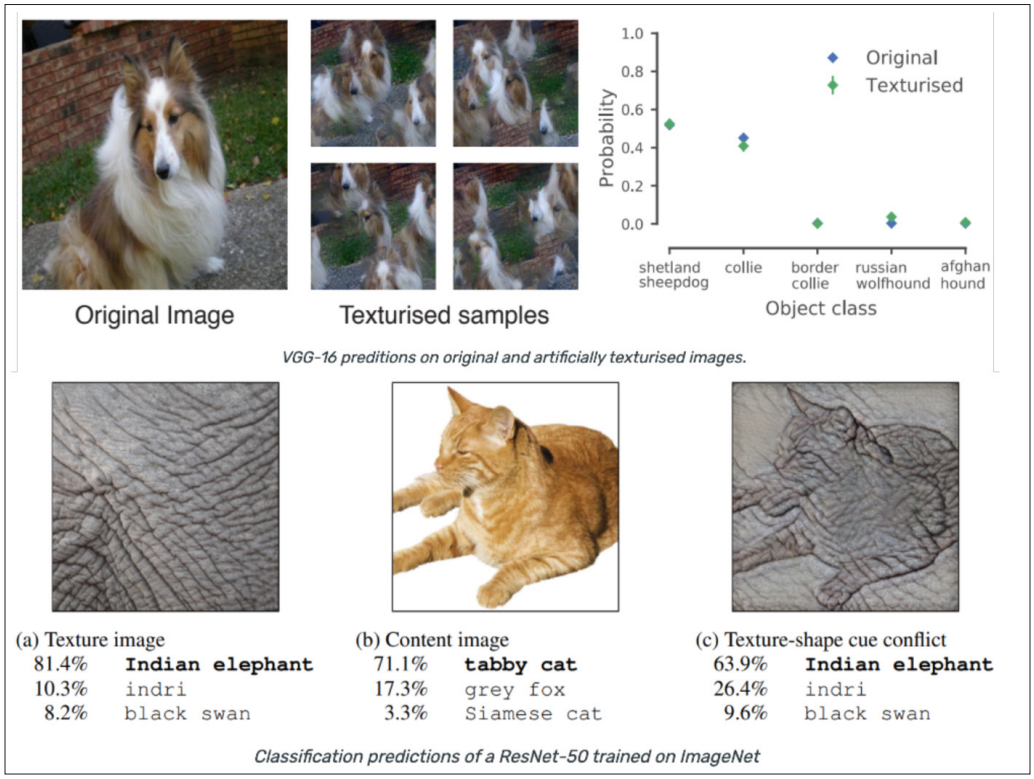
□ 김승룡 / 고려대학교

요약

본 고에서는 최근 컴퓨터 비전 분야에서 가장 활발히 연구되고 있는 분야 중에 하나인 자가 학습(Self-supervised Learning) 기술의 동향과 향후 방향성에 대해서 논의한다. 컴퓨터 비전 분야에서의 자가 학습 기술은 최근에 Contrastive Learning 기법을 활용하여 활발하게 연구되고 있는데, 이를 위한 좋은 Positive와 Negative를 어떻게 추출할까에 대한 고민으로 수많은 연구들이 진행되어 왔다. 본 고에서는 이러한 방향성에서 대표적인 몇 가지의 방법론에 대해서 논의하고 이의 한계점을 언급하며 컴퓨터 비전 분야에서 자가 학습 기법이 가야할 방향성에 대해서 논의하고자 한다.

I. 서 론

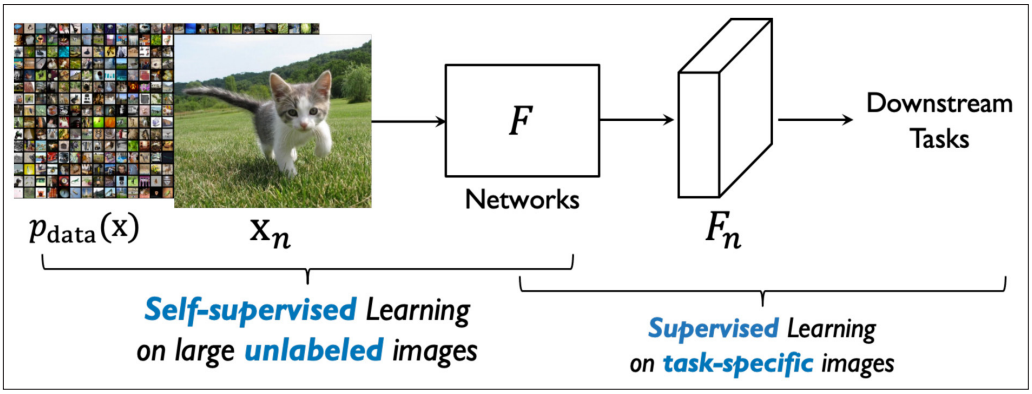
현대의 인공지능 및 컴퓨터 비전의 현저한 발전은 ImageNet을 필두로 한 대규모(Large-Scale)의 라벨이 존재하는(예: 영상 클래스 라벨 등) 데이터의 발현과 이를 활용한 딥 뉴럴 네트워크(Deep Neural Networks)의 발견이 시발점이 되었다. 이러한 딥러닝과 교사 학습(Supervised Learning)의 조합은 현대의 인공지능의 발전에 밑거름이 되었고, 수많은 분야에서 이러한 기법들이 활용되어 현저한 성능 향상을 달성하였다. 이러한 교사 학습 기반의 인공지능 기술들은 단순히 특정 테스크(Task) 또는 도메인(Domain)의 데이터와 라벨을 활용하여 문제를 푸는 것 뿐만 아니라, 특정 테스크 또는 도메인에서 학습된 네트워크를 다른 테스크 또는 도메인으로 전달(Transfer)하는 데 많이 사용되고 있다. 예를 들어 대부분의 컴퓨터 비전 기술을 활용한 솔루션은 네트워크를 학습하는 과정에서 스크래치(Scratch)에서 시작하는 것이 아니라, ImageNet에서 기 학습된 모델을 활용하여 Fine-tuning하는 게 일반적이었다. 하지만 이러한 교사 학습 기반의 인공지능 기술들은 대규모의 데이터셋과 이의 라벨링(Labeling)을 필요로 한다는 측면에서 매우 도전을 가지고 있으며, 특정한 도메인(예: 의학 도메인 등)은 이러한 라벨링조차 대규모로 얻



<그림 1> 영상 분류에서 교사 학습 기법의 한계 예시([1] 논문에서 발췌)

는 게 불가능한 경우도 존재한다. 또한 이러한 대규모의 데이터셋과 라벨링을 획득한다고 해도 라벨링 자체에서 포함되는 바이어스(Bias)가 네트워크 학습에 영향을 미칠 수 밖에 없다. 예를 들어 영상 분류(Image Classification)의 경우 특정 클래스의 영상을 분류할 때 영상에 존재하는 모든 정보를 포괄적으로 쓰는 게 아닌 텍스쳐(Texture) 등만을 보고 영상 분류를 수행하는 네트워크가 배워지는 경우가 있고 이는 네트워크의 일반화 성능(Generalization Ability)을 제한한다(〈그림 1〉)[1,2].

이러한 한계를 극복하기 위하여 최근 인공지능 기술들은 라벨링 정보를 사용하지 않고 영상 정보만을 활용하여 학습하는 자가 학습(Self-supervised Learning) 기법에 널리 연구되고 있다[3,11,14]. 이러한 자가 학습 기법은 값비싼 어노테이션(Annotation) 과정을 건너뛰고 학습을 진행할 수 있다는 측면에서 매우 매력적인 연구 분야이고, 수없이 존재하는 영상 또는 비디오 자체를 학습 데이터로 활용할 수 있다는 측면에서 매우 의미있는 방향성을 보여준다(예: 300시간의 비디오가 매 1분마다 YouTube에 업로드된다고 알려져 있다.). 이러한 자가 학습 기법은 단순히 연구적으로 제시된 방향성이 아니라 실제 신생아 또는 어린아이가 세상을 인식하고 인지하는 과정 또한 라벨링이 제한된 자가 학습이라는 측면에서 앞으로의 인공지능이 나아가야 할 방향성임에는 틀림이 없다. 또한 이러한 자가 학습 기법은 특징 표현자 학습(Representation Learning)에도 사용될



<그림 2> 자가 학습(Self-supervised Learning)과 이를 활용한 튜닝 기법의 예시

수 있어 대규모의 이미지 또는 영상 데이터를 활용하여 딥러닝 네트워크를 학습하여 특정 테스크 또는 도메인에 아주 간단한 튜닝만을 통하여 성능을 내는 데 활용될 수 있다는 측면에서 의미있는 방향성이라 할 수 있다(〈그림 2〉).

사실 이러한 자가 학습 기술에 대한 필요성은 딥러닝의 혁신(Revolution) 전부터 많이 고민되어 온 문제이고, 딥러닝 시대(2012년 AlexNet의 발견 이후)에도 수많은 연구자들이 고민해 온 분야이다. 하지만 최근까지 이러한 자가 학습 기법들이 연구의 주요한 방향성으로 각광받지 못한 가장 큰 이유는 성능 자체의 제약 때문이었다. 수많은 자가 학습 기법들이 나왔지만 그 당시까지만 해도 교사 학습 기법의 성능을 따라가지 못해서 이기도 하다. 하지만 최근에 Contrastive Learning 기법을 필두로 한 수많은 기법들이 교사 학습 기법의 성능을 뛰어넘으며 자가 학습 기법에 대한 방향성의 정당성을 보여주고 있다[14,15,16]. 본 고에서는 인공지능 및 컴퓨터 비전 분야에서 이러한 자가 학습 기법에 대한 전통적인 방향성에 대한 고찰에서 시작하여, 최근 자가 학습 기법들의 방법론 및 성능 분석, 그리고 이들의 한계점을 분석해 보고자 한다.

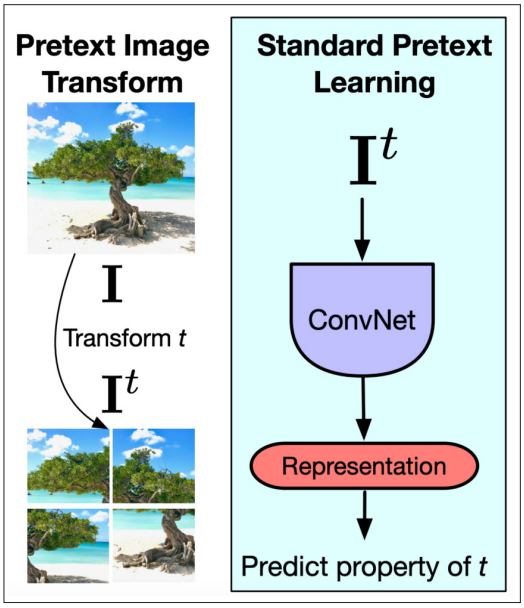
구체적으로 II장에서는 최신 자가 학습 기반 인공지능 기술 동향에 대해서 소개하고, III장에서는 이러한 기술들의 성능과 한계점에 대해서 논의할 예정이며, IV장에서는 결론을 맺으며 본 고를 마치려 한다.

II. 최신 자가 학습 기반 인공지능 기술 동향

컴퓨터 비전(Computer Vision) 분야에서 자가 학습 기반 인공지능 기술이 크게 대두된 것은 비교적 최근이지만 자연어 처리(Natural Language Processing) 분야에서는 자가 학습 기반 인공지능 기술이 빠르게 발전되어 왔고 정형화되어 왔다[3]. 대표적으로 BERT[3]라고 불리우는 모델은 Pre-training of Deep Bidirectional Transformers for Language Understanding 연구로써 Missing Word Prediction과 Next Sentence Prediction을 자가 학습을 위한 손실 함수로 삼아 학습을 수행하였고, 교사 학습 기법에 비해 매우 높은 성능을 보여주며, 자연어 처리 분야에서는 자가 학습 기법을 활용하여 학습된 모델을 쓰는 것이 정형화되어 왔다. 하지만 이에 비해 컴퓨터 비전 분야는 영상이 가지고 있는 특성들이 자연어와 많이 다르기 때문에 자가 학습 기술

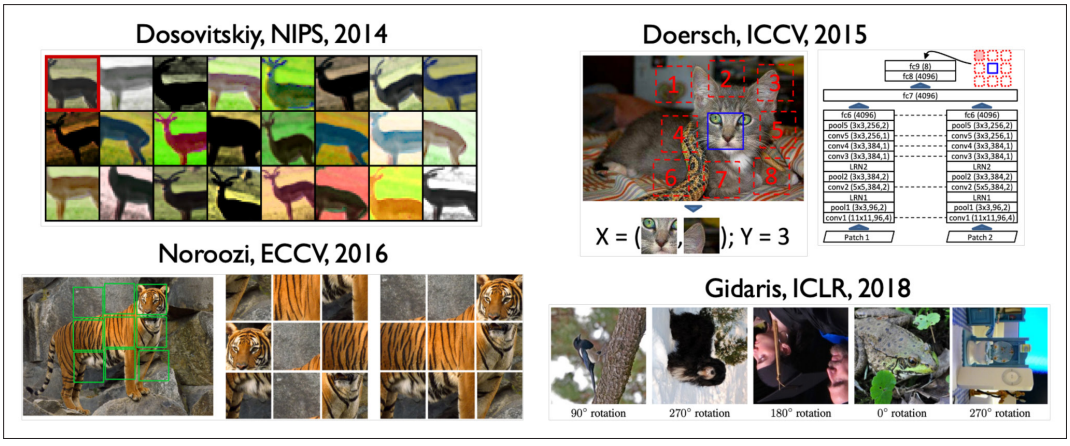
은 비교적 더디게 진행되어 왔다.

컴퓨터 비전 분야에서는 자가 학습 기술은 Pretext Task를 어떻게 잘 정립하고 이를 활용하여 어떻게 좋은 특징 표현자를 잘 학습할 수 있는가에 관한 연구이다(〈그림 3〉). 영상에 제안된 Pretext를 활용하여 왜곡(Deformation)을 넣어 주고 딥러닝 네트워크가 다시 한 번 그 왜곡을 잘 예측하는 특징 표현자를 학습하도록 하는 게 가장 일반적인 Pretext Task 기반의 자가 학습 기법이다[4,5,6]. 이러한 Pretext를 어떻게 정의하느냐에 따라서 수많은 연구들이 진행되어 왔으며, 각자 본인들이 제안하는 Pretext에 따라서 네트워크 학습을 진행하여 성능의 향상을 확인하는 식으로 연구가 되어 왔다.



<그림 3> 전통적인 Pretext 기반의 자가 학습 기술의 예시([4] 논문에서 발췌)

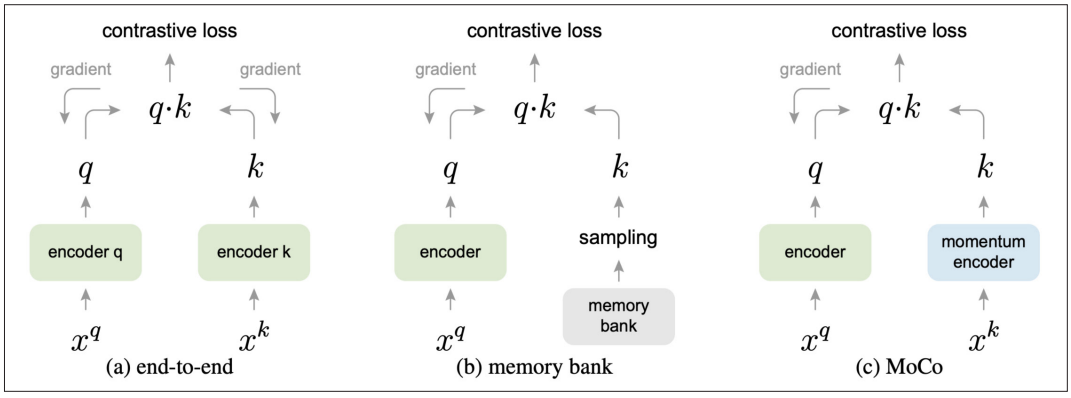
예를 들어 Pretext를 어떻게 정의하느냐에 따라서 방법론이 달라지는데(〈그림 4〉), 단일 영상에 다양한 데이터 증분을 적용한 후 같은 특징 표현자가 생성되도록 수행하는 연구가 있었고[5], 영상에서 다양한 패치(Patch)들을 샘플링한 후 그 패치들의 기하학적 관계를 예측하도록 설계한 연구가 있었으며[6], 이러한 패치들을 잘 섞어서 직쏘(Zigsaw) 퍼즐을 예측하는 식으로도 설계가 되었다[7]. 또한 영상에 임의의 회전(Rotation) 왜곡을 적용하여 해당하는 회전 왜곡을 예측하도록 학습하여 특징 표현자를 학습하는 기술도 제안되었다. 또한 영상의 컬러를 없앤 후 다시 컬러를 복원한다든지[9], 영상에서 임의로 패치를 삭제한 후 다시 패치를 복원하는 식의 Pretext도 연구가 진행되었다[10]. 이러한 기술들은 라벨을 활용하지 않고 네트워크를 학습할 수



<그림 4> 전통적인 Pretext 기반의 자가 학습 기술의 예시([5,6,7,8] 논문에서 발췌)

있다는 측면에서 매우 매력적인 방향성을 보여주었으나 교사 학습 기법들과의 성능 차이로 인하여 크게 성공을 거두지는 못하였고 컴퓨터 비전의 큰 방향성은 이때까지만 해도 교사 학습 기법으로 학습된 네트워크를 활용하는데 초점이 맞추어져 있었다.

이러한 연구의 방향은 Contrastive Learning을 통한 자가 학습 기법의 발현을 통하여 큰 성공을 거두었는데, Contrastive Learning을 활용한 다양한 자가 학습 기법들이 소개되기 시작하며 여러 가지 테스크에서 교사 학습 기법을 뛰어넘는 성능을 보여주었다. 이러한 연구의 시작은 SimCLR[11]이라고 할 수 있는데, 단일 영상에서 두 가지의 데이터 증분(Data Augmentation)을 활용하여 두 가지의 다른 영상을 만들고 이를 Positive라고 가정하고 다른 영상들을 Negative라고 가정한 후, InfoNCE[11] 기반 손실 함수를 활용하여 네트워크를 학습하는 기법을 제안하였다. 이러한 SimCLR은 어떠한 Supervision도 사용하지 않고 데이터 증분만을 활용하여 Contrastive Learning을 통한 자가 학습을 가능케 하였다는 데 큰 의미가 있었고 실제 많은 테스크에서도 이전 방법론들의 성능을 크게 뛰어넘으며 가능성을 보여주었다. 이러한 InfoNCE[12,13] 기반의 Contrastive Learning은 충분히 많은 Negative 샘플들이 있어야 성능이 보장된다는 한계를 가지고 있었는데, SimCLR 논문에서도 이를 해결하기 위하여 배치(Batch) 내의 영상을 약 4,096개 이상 사용하여 이들의 조합을 가지고 Negative 샘플을 만들어 학습을 진행하였다[11]. 이러한 설계 방안은 GPU의 용량 및 속도에 의존도가 높아 사용이 제한적이었고, 이후의 연구는 이러한 한계를 어떻게 뛰어넘을지에 초점을 맞추었다. SimCLR 연구와 비슷하게 메모리 뱅크(Memory Bank) 개념을 사용하여 Negative를 생성하고 이를 활용하여 Contrastive Learning을 수행하는데 이때 메모리 뱅크 쪽으로는 손실 함수에 그래디언트(Gradient)를 흘려주지 않는 방법으로 Negative 샘플링을 위한 큰 스케일의 배치 사이즈의 문제를 해결하려 하였는데, 메모리 뱅크 안에 존재하는 특징 표현자가 현재의 업데이트된 특징 표현자와 크게 상이하여, 너무 쉬운 Negative로 고려되었고, 따라서 제한적인 성능을 보여주었다. 이를 해결하기 위하여 MOCO[14]라는 방법론에서는 Momentum Encoder 개념을 제한하여 메모리 뱅크에 저장된 영상을 Momentum Encoder를 활용하여 특징 표현자를 추출하고 이를 활용하여 Contrastive Learning을 하



<그림 5> Contrastive Learning 기반 자가 학습 기술의 예시([11,14] 논문에서 발췌)

는 방향으로 확장되었다. 이러한 기법은 Momentum Encoder가 천천히 기존 Encoder를 따라가게 하여 안정적인 학습이 가능해졌다. 〈그림 5〉는 이러한 초창기 연구의 도식도이다.

이러한 연구들은 향후에 Negative Sample을 완전히 사용하지 않는 방향들로 다양하게 확장이 되었는데, Bootstrap Your Own Latent(BYOL)[15], SwAV[16], SimSiam[17] 같은 방법들은 Negative Sample들을 전혀 사용하지 않고 Positive Sample들만을 효과적으로 활용하여 자가 학습 성능을 극대화하였다. 또한 최근에는 다양한 컴퓨터 비전 분야에서 Transformer들을 활용하여 성능을 극대화하고 있는데, 이런 Vision Transformer에도 자가 학습을 이용해서 학습하는 다양한 방법론들, Self-supervised Vision Transformer(SiT)[18], DINO[19], BEIT[20], Masked Autoencoder[21]와 같은 기술들이 활발히 연구되어 오고 있다.

III. 기존 방법들의 한계점 및 향후 연구 방향

기존의 자가 학습 기법들은 Contrastive Learning을 활용하여 그 성능이 극대화되었고, Vision Transformer와 같은 다양한 구조(Architecture)에서도 좋은 성능을 보여주고 있다. 하지만 이러한 자가 학습 기법들은 영상 레벨(Image-level)에서는 활발히 연구되어 왔으나 픽셀 레벨(Pixel-level)에서는 제한적으로 연구가 되어왔다는 한계를 가진다. 최근 이러한 픽셀 레벨의 특징 표현자도 자가 학습 기법을 활용하여 잘 학습하려는 시도들이 있었지만 성능이 제한적이었기에, 이를 극복할 수 있는 방법론들에 대한 고민이 필요하다. 또한 최근에 많은 적용 분야에서 컴퓨터 비전과 자연어 처리를 동시에 활용하려는 움직임이 있는데, 이러한 영상과 자연어를 동시에 활용하여 좋은 특징 표현자를 학습하려는 연구는 제한적으로 시도되어서 이를 해결하는 연구 또한 매우 중요해질 것이다.

IV. 결 론

본 고에서는 최근 인공지능 분야에서 가장 각광받는 기법 중에 하나인 자가 학습 기법에 대한 최근 연구 동향에 대해서 논의하였다. 컴퓨터 비전 분야의 자가 학습 기법은 Contrastive Learning을 활용한 솔루션들이 높은 성능을 나타내며 성공을 거두었는데, 영상만을 활용하여 어떻게 Positive와 Negative Sample들을 찾아내는지에 대한 연구가 주를 이루어 왔다. 또한 이러한 기존의 자가 학습 기법에 대한 한계점와 향후 방향성에 대하여 논의를 진행하였다. 이러한 자가 학습 기법은 그 잠재력과 활용성이 무궁무진하므로 이에 대한 연구 및 논의는 계속 필요할 것이다.

참 고 문 헌

[1] L.A. Gatys et al., Texture and art with deep neural networks, Neurobiology, 2017
[2] R. Geirhos et al., ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, ICLR 2019
[3] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019
[4] Misra and Maaten, Self-Supervised Learning of Pretext-Invariant Representations, ArXiv, 2020
[5] Dosovitskiy et al., Discriminative Unsupervised Feaature Learning with Exemplar Convolutional Neural Networks, NIPS, 2014
[6] Doersch et al., Unsupervised Visual Representation Learning by Context Prediction, ICCV, 2015
[7] Norrozi et al., Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, ECCV, 2016
[8] Gidaris et al., Unsupervised Representation Learning by Predicting Image Rotations, ICLR, 2018
[9] Zhang et al., Colorful Image Colorization, ECCV, 2016
[10] Pathak et al., Context Encoders: Feature Learning by Inpainting, CVPR, 2016
[11] Chen et al., A Simple Framework for Contrastive Learning of Visual Representations, ICML, 2020
[12] Z. Wu et al., Unsupervised Feature Learning via Non-Parametric Instance Discrimination, CVPR, 2018
[13] I. Misra et al., Self-Supervised Learning of Pretext-Invariant Representations, CVPR, 2020
[14] K. He et al., Momentum Contrast for Unsupervised Visual Representation Learning, CVPR, 2020
[15] J. B. Grill et al., Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, NeurIPS, 2020
[16] M. Caron et al., SwAV: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, NeurIPS, 2020
[17] X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR, 2021
[18] S. Atito et al., SiT: Self-Supervised Vision Transformer, ArXiv, 2021
[19] M. Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ArXiv, 2021
[20] H. Bao et al., BEIT: BERT Pre-Training of Image Transformers, ICCV, 2021
[21] K. He et al., Masked Autoencoders Are Scalable Vision Learners, ICCV, 2021

필 자 소 개

김 승 룡

- 2012년 : 연세대학교 전기전자공학과 학사
- 2018년 : 연세대학교 전기전자공학과 박사
- 2018년 ~ 2019년 : 연세대학교 전기전자공학과 박사후 연구원
- 2019년 ~ 2020년 : 스위스 EPFL 박사후 연구원
- 2020년 ~ 현재 : 고려대학교 컴퓨터학과 조교수
- ORCID : https://orcid.org/0000-0003-2927-6273
- 주관심분야 : 컴퓨터 비전, 기계학습, 인공지능 등