Self-supervised Learning on Billion unlabeled Image data

CUAI 4 기 skt ai fellowship 팀
김민지(응용통계학), 신재현(컴퓨터공학), 정현희(응용통계학)

[요약] Computer Vision 분야에서의 핵심은 적용하고자 하는 input set 에 대한 좋은 representation 을 뽑는 작업이라고 할 수 있다. 우리는 unlabeled 데이터를 활용하여 비교적 적은 labeled 데이터 없이도 optimized parameter 를 찾을 수 있는 pseudo labeling 기법과 self-supervised learning 을 통해 사용하고자 하는 industry 에 맞는 representation 을 학습하고 이를 upstream weight 으로 사용하여 selected ImageNet, CIFAR10, CIFAR100, VERI-Wild, COCO 데이터셋에 대해 less labeled set 으로 성능을 높일 수 있으며 학습속도 또한 빠른 방법임을 실험적으로 증명했다. 더 나아가, 이후에는 SK Telecom 에서의 real-world 데이터를 사용하여 실제 데이터에도 잘 적용됨을 증명하고자 한다. 또한 metric learning 방법을 접목하여 더욱 다양한 분야에서 data agnostic 하게 활용할 수 있도록 추가 실험을 진행할 계획에 있다.

1. 서 론

최근 딥러닝 모델은 Human Intelligence 에 견줄만큼 큰 발전을 이뤄내고 있다. 하지만 딥러닝의 큰 단점은 billions of labeled 데이터가 필요하다는 점과 학습속도가 오래 걸린다는 점이다. 최근까지 딥러닝에서의 지배적인 패러다임은 supervised 방법이다. supervised 방법은 정의된 라벨에 따라 model parameter 을 업데이트하며 학습하는 방식이다.

self-supervised 방법은 supervised 방법과 unsupervised 방법의 중간 형태로 볼 수 있는 방법이다. unsupervised 방법은 지정된 labeled 값 없이 embedding 된 벡터값들의 distance 등을 구하여 clustering, dimensional reduction 과 같은 방법을 사용하여 문제를 푸는 방법이다. self-supervision 이란 이미지의 특징을 추출하는 모델이 있다고 할때 인풋 데이터의 한 부분이 다른 부분의 supervision 역할을 하게 합니다.

Computer Vision 분야에서 self-supervised learning 을 적용하기 쉽지 않은 이유는 data 가 discrete 하지 않고 continuous 하기 때문에, missing part 를 찾는 것이 쉽지 않기 때문이다. Missing part 는 video 에서의 frame, image 에서의 patch 를 말한다.

이런 한계점이 있음에도 self-supervised learning 을 Computer Vision 분야에 활용하기 위한 연구가 지속적으로 있었다. NeurIPS 2020 에서 그간 등장하지 않았던 self-supervised learning 의 주요 방법 중 하나인 대조학습이 주요 키워드로 부상했다.

최근에는 contrastive learning 방식들이 제안되고 있다. 특히 Simple Framework for Contrastive Learning of visual Representation(SimCLR)을 시작으로 negative sample 없이 Teacher-Student model 을 사용하여 contrastive learning 을 진행하는 Bootstrap Your Own Latent(BYOL), Siamese Network 를 사용하여 augmented 값과 anchor 값의 차이를 contrastive loss 를 사용하여 계산하는 Siamese representation(SimSiam)과 같은 방법들이 두각을 나타내고 있다.

우리는 제안된 self-supervised 방법들을 real world set 에 적용하여 실험한후 배포하여 모델을 활용하고자 한다. 특히 적용하고자 하는 분야인 vehicle type classification 과 vehicle model detection 은 제조사와 모델명 등을 하나하나 label 을 달아주는 것이 어려운 만큼, large scale 의 unlabeled data 는 있지만 제대로 활용을 못하고 있는 대표적인 알고리즘이다. Vehicle 의 경우 한국형 데이터를 얻기 힘들고, 차량의 경우 개인정보 문제까지 있다. 따라서 먼저 self-supervised learning 을 적용한 vehicle classification 모델을 Vehicle Re-Identification task 에 먼저 적용하여 성능을 확인해보았다. 또한 certainty 기반의 Human In the Loop Learning (HILL) 프레임워크를 구축하여 성능을 더욱 높여 SK Telecom 에서 배포될 예정이다.

2. 본 론

1) Contrastive Learning : Siamese Representation

Self-supervised Learning 에서 활용되는 contrastive learning 의 핵심은 이미지 x 와 y 가 anchor 이미지를 기준으로 augmented 된 이미지라고 할 때, low energy 를 갖도록 visual representation 학습이 이루어지는 방식이다. Siamese Network 는 두 개의 입력에 대해 독립적으로 두 개의 합성곱 신경망을 실행한 뒤 비교하는 아이디어입니다. 거리 기반 학습 방식의 대표적인 모델 중 하나답게 Siamese Neural Network 는 클래스가 같은 샘플 사이의 거리는 가깝도록, 클래스가 다른 샘플 사이의 거리는 멀어지도록 특징을 추출하는 목적식을 구성하게 된다 . 또한 두 augmented image 의 representation 의 similarity 를 기반으로 학습하므로 Few-shot learning 에도 적용이 가능하다.

Energy Based Model 을 학습할 때 가장 큰 문제점은 (x,y)가 완전히 다를 때도 동일한 입력이라고 판단하여 low energy 를 갖게 되는 representation collapse 가 발생할 수 있다는 점이다. SimCLR 과 MoCo 모두 Contrastive learning 을 활용하는 Siamese network 기반이기에 representation colapse 를 겪고, 이것을 해결하기 위한 기법으로 SimSiam 이 제시되기도 했다.

2) Pseudo-labeling method

Consistency regularization 이 가지는 단점을 보완하는 method 로 semi-supervised learning, 그 중에서도 pseudo-labeling 에 대해 고려하게 되었다.

*1)*에서 논한 바와 같이, consistency regularization 방식을 따르게 되었을 때의 명확한 단점이 존재한다. (representation collapsing 이 일어날 수 있다는 점) 또한 동일한 이미지를 augment 한 것은 유사도가 높다고 학습하고, 각각 다른 이미지를 augment 한 경우 유사도가 낮다고 학습하는 개념이기 때문에 domain-specific data augmentations 에 의존할 수밖에 없다는 문제도 존재한다. Billion unlabeled image data 를 이용하는 본 연구에서, 데이터의 모든 modality 를 커버할 수 있는 augmented data 를 다 만들어내는 것은 상당히 어려울 것이다. 최대한 많은 modality 를 커버할 수 있게 노력한다고 했을 때, computing resource 의 한계로 testing 자체에 상당히 많은 시간이 소요될 것이라는 판단을 내렸다.

이러한 제약이 없는 방법이 바로 pseudo-labeling 이나, consistency regularization 방식에 비해 성능이 비교적 떨어진다는 단점이 있다. pseudo-labeling 이 비교적 낮은 성능을 보이는 것은, poorly calibrated model 들에서 오는 erroneous high confidence prediction 들 때문이다. 이는 잘못된 pseudo-label 들의 생성으로 이어지고, 결국 noisy training 에 의해 성능이 떨어지게 된다.

training process 로 들어가는 noise 를 현격히 줄여 pseudo labeling accuracy 를 높이고, negative pseudo label 의 생성을 허용하여 pseudo-labeling process 를 일반화하여 성능에서의 문제까지 해결한 framework 가 바로 UPS(Uncertainty-aware Pseudo-label Selection) framework 이다.

Consistency regularization 방법이 가지는 단점을 보완하면서 전통적인 pseudo-labeling 기법들이 가지는 성능 문제까지 해결했기에 UPS 의 효과 검증을 거쳐 task 에 실제로 적용해보고자 했다. 이를 위해 [1]에서 experiment 가 이루어진 benchmarking dataset 은 물론이고 그렇지 않은 dataset 에 대해서도 experiment 를 진행해보았다. **Figure 1.**은 진행한 데이터에 대해 간단히 분석을 해본 결과이다.

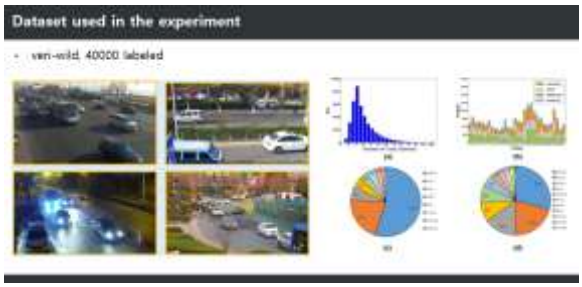

Figure 1. VERI-Wild 데이터셋 1

실험 결과는 다음과 같다.

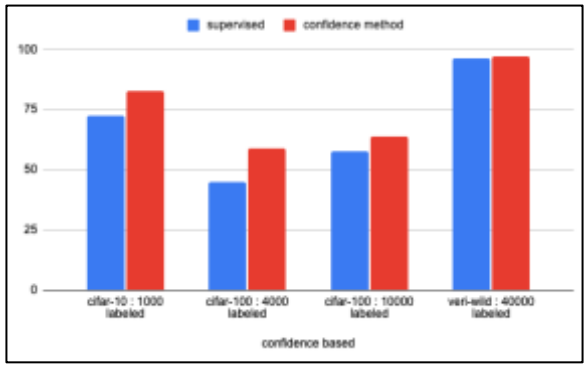

Figure 2. confidence based method 결과 1

Confdience based model 은 기존의 pseudo labeling 에서 많이 쓰던 방법으로, dataset 을 한번 파싱한 후에 retraining 시킨 결과이다.

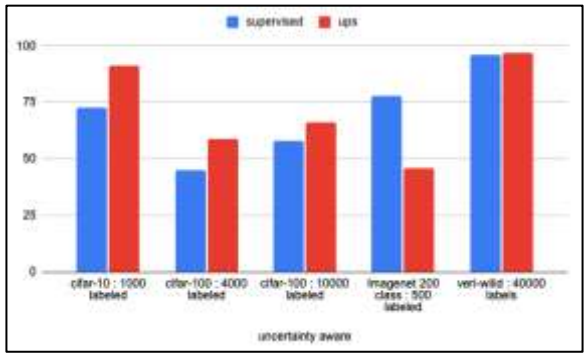

Figure 3. uncertainty aware 결과 1

Figure 3. 의 uncertainty aware 방법이 [1]에서 cifar-10 및 cifar-100 에 대하여 experiment 를 진행한(실제로 그러한 성능이 나와주는지 다시 test 진행해본 바 있음), 그리고 본 논문에서 ImageNet 및 VeRI-WILD 에 추가적인 experiment 를 진행한 UPS 방법에 해당한다.

거의 대부분의 상황에서 빨간색 바와 파란색 바의 차이가 confidence based 의 경우보다 uncertainty aware 의 경우에 더 큼을 알 수 있다. 예외적으로 ImageNet 의 경우 labeled set 으로만 학습한 결과, 즉 파란색 바가 더 높게 나타났다. labeled set 개수, train model architecture, certainty thresholding 값 등에 대한 최적화가 되지 않은 채로 training 을 진행했기 때문인 것으로 판단되며 추가 실험 진행 예정에 있다.

그리고 open dataset 중 VeRI-WILD dataset 의 경우, ups 와 certainty based 방법의 차이가 크지 않았다. 이 결과를 통해서 well-calibrated 된 상황에서는 둘 간의 큰 차이가 없다는 결론을 내릴 수 있었다.

3) State of Art in Self-supervised Learning

Consistency regularization, Generative method, Transformer based method 등 다양한 Self-supervision approach 들 중 Consistency Regularization 쪽에 집중을 하기로 했다. Generative method 는 다른 approaches 에 비해 성능이 떨어지는 동시에 차종 분류 task 에 최적화된 method 는 아니라는 판단을 했고, transformer based 는 비교적 검증이 덜 이루어져 있기 때문에 검증이 더 잘 되어 있는 method 에 집중하는 편이 좋겠다는 결론을 내렸다. Consistency Regularization method 들 중 BYOL 과 SimSiam 을 실험 대상으로 선정한 이유는 다음과 같다.

BYOL 의 경우 negative pairs 없이 당시 SOTA 를 능가(linear evaluation protocol 이용)했을 뿐만 아니라 batch size 나 image augmentation 의 변경에 대해 resilient 하다는 장점이 있기 때문에 our method 확립 시 참고할 만한 요소가 많다고 판단했다.

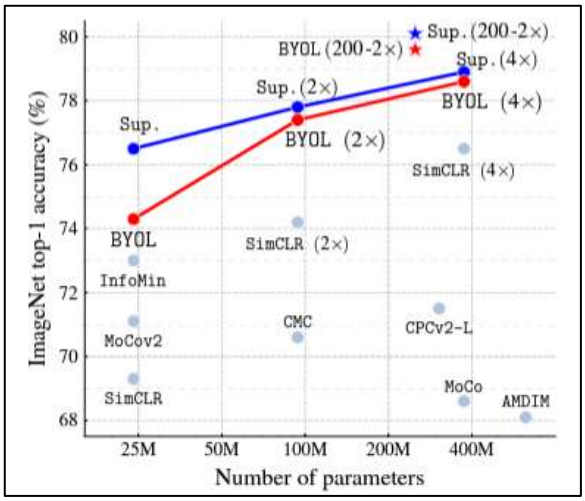

Figure 4. ImageNet 에서의 BYOL 성능 1

SimSiam, 즉 Simple Siamese 는 consistency regularization method 에서 발생하는 collapsing solution 문제를 기존 방법들과는 다른, 새로운 방식으로 해결한다. SimCLR 나 BYOL 등의 다른 방법에서는 모든 출력값이 상수로 무너지는 collapsing solution 문제 해결을 위해 negative sample pairs, large batches, momentum encoders 등을 흔히 이용한다. 그러나 이런 해결책들은 모두 컴퓨팅 리소스를 많이 잡아먹기 때문에 마냥 좋은 해결책이라고 보기에는 어려움이 있다. SimSiam 은 그런 방식들 대신 same encoder, prediction MLP, stop-gradient 의 세 가지 대안을 제시한다. 간단하면서, collapsing solution 문제도 획기적인 방식으로 해결했기 때문에 선정하게 되었다.

3. 결 론

여러 benchmark set 에 대해 UPS method 를 test 해본 결과, 단순 supervised 에 비해 pseudo-labeling(confidence-based)이 가지는 효과, 그리고 confidence-based 에 비해 uncertainty aware pseudo-label selection 이 가지는 효과가 존재함을 확인하였다. 또한 well-calibrated 상황에서는 ups 와 confidence based method 간에 큰 차이가 없다는 결론을 내릴 수 있었다. 그러나 기대한 만큼의 결과가 나오지 않았기 때문에 hyperparameter tuning 을 통해 dataset 에 최적화된 framework 로의 수정이 이루어질 필요가 있다. 조정 후 다시 실험이 이루어질 계획이며 certainty measuring 하는 metric 을 더 추가하여 selected certainty

image 를 더 strict 하게 가려내볼 계획이다.

Table 1. Comparison on Imagent linear classifcation on Resnet-50

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

Table 1. 을 보면 100 epoch 까지만 돌렸을 때 SimSiam 이 가장 좋은 성능을 내지만, 200 부터는 BYOL 이 가장 좋은 성능을 보여주고 있음을 알 수 있다. Epoch 수를 늘리면 BYOL 이 우세하기 때문에 BYOL 을 완전히 대체할 수 있다고 주장할 수 없다는 점에서 아쉬움이 있다. 물론 현실적으로 본 연구에 투입 가능한 GPU resource 를 고려하면 SimSiam 이 강점을 가지기에 앞으로의 experiments 는 SimSiam 을 main 으로 하여 진행될 예정이다.

SimSiam 은 negative sample 이나 momentum encoder 를 없애고 stop gradient 를 통해 collapsing solution 문제를 해결할 수 있었다는 것이 empirical 하게 증명되었으나 수학적으로 증명된 바가 없다. 수학적, 이론적인 증명이 추가된다면 추후 성능 향상에 근본적으로 도움을 줄 수 있을 것으로 보인다.

차종 detection(Vehicle type detection)의 경우 세단과 경차 등등 차종에 따라서 주차요금을 다르게 매기는 서비스에 사용될 수 있다. 차량 모델 detection(Vehicle model detection)의 경우 무인 주차장 입출차 통계를 내거나, 백화점에서 구매력이 높은 고객을 파악하는 용도 등으로 사용될 수 있을 것이다.

Detection 은 classification 과정을 포함하며, classification 문제는 특정 데이터에서 좋은 성능을 보였을 때 다른 데이터에서도 괜찮은 성능을 보일 가능성이 높기 때문에 먼저 차종 및 차량 모델 detection 에 대해서 좋은 성능을 내는 our method 를 찾는 것이 급선무이다. 차종과 차량 모델을 잘 detect 할 수 있게 되면 추후 domain adaptation 을 통해 다른 데이터에도 적용해볼 예정이다.

참고 문헌

[1] "Energy Base Model description". Available: https://atcold.github.io/pytorch-Deep-Learning/ko/week07/07-1/

[2] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". Computer Vision and Pattern Recognition (CVPR) 2020, Nov. 2020

[3] Chen, Ting, Simon Kornblith, Mohammad Norouzi and Geoffrey E. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations", International Conference on Machine Learning (ICML), Feb. 2020

[4] . He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nov. 2019

[5] "VERI-Wild Github Link" . Available: https://github.com/PKU-IMRE/VERI-Wild

[6] Rizve, M. N., Kevin Duarte, Y. Rawat and M. Shah. "In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning". International Conference on Learning Representations (ICLR) 2021, Jan. 2021

[7] Grill, Jean-Bastien et al. "Bootstrap your own latent: A new approach to self-supervised learning", NeuIPS 2020 , Jun. 2020