자연 영상에 대한 Naive Convolutional Auto Encoder의 특징 추출 성능에 관한 연구

*이성주 **조남익
서울대학교 전기정보공학부 뉴미디어통신공동연구소
thomas11809@snu.ac.kr nicho@snu.ac.kr

A Study on Feature Extraction Performance of Naive Convolutional Auto Encoder to Natural Images

*Lee, Sung Ju **Cho, Nam Ik
Department of ECE, INMC, Seoul National University

요약

최근 영상 군집화 분야는 딥러닝 모델에게 Self-supervision을 주거나 unlabeled 영상에 유사-레이블을 주는 방식으로 연구되고 있다. 또한, 고차원 컬러 자연 영상에 대해 잘 압축된 특징 벡터를 추출하는 것은 군집화에 있어 중요한 기준이 된다. 본 연구에서는 자연 영상에 대한 Convolutional Auto Encoder의 특징 추출 성능을 평가하기 위해 설계한 실험 방법을 소개한다. 특히 모델의 특징 추출 능력을 순수하게 확인하기 위하여 Self-supervision 및 유사-레이블을 제공하지 않은 채 Naive한 모델의 결과를 분석할 것이다. 먼저 실험을 위해 설계된 4가지 비지도학습 모델의 복원 결과를 통해 모델별 학습 정도를 확인한다. 그리고 비지도 모델이 다량의 unlabeled 영상으로 학습되어도 더 적은 labeled 데이터로 학습된 지도학습 모델의 특징 추출 성능에 못 미침을 특징 벡터의 군집화 및 분류 실험 결과를 통해 확인한다. 또한, 지도학습 모델에 데이터셋 간 교차 학습을 수행하여 출력된 특징 벡터의 군집화 및 분류 성능도 확인한다.
*부록을 포함한 전체 논문은 저자에게 요청하시오. (이성주, thomas11809@snu.ac.kr)

1. 서론

지도학습(supervised learning)을 통한 딥러닝 영상 분류는 시간이 지남에 따라 점점 완전에 가까운 성능을 보여주고 있다. 이와 같이 레이블이 있는 데이터셋에 대한 분류 정확도가 100%로 포화 되어 가는 한편, 현실에선 레이블이 없는 영상을 학습시켜야 하는 상황들이 많이 존재한다. 즉, 비지도학습(unsupervised learning)을 통한 영상 군집화(image clustering)의 필요성은 나날이 중요해지고 있고, 해당 문제에 딥러닝을 사용하려는 시도가 늘어나고 있다.

최근 영상 군집화 분야는 딥러닝 모델에게 Self-supervision을 주거나 unlabeled 영상에 유사-레이블을 주는 방식으로 연구되고 있다. DEC[1]는 사전 학습된 Stacked Auto Encoder를 이용하여 입력 영상을 잠재 공간(latent space)으로 임베딩(embedding)한 특징 벡터를 사용한다. 이때 특징 벡터에 K-means 알고리즘을 적용하여 군집(cluster)의 중심으로부터 soft assignment를 얻어내고, 이를 통해 목표하는 유사-진 분포를 만드는 방식으로 신경망 모델에게 Self-supervision을 적용한다. DCEC[2]는 DEC와 같은 방법을 사용하되 특징을 추출하기 위해 완전연결계층 대신 Covolutional Auto Encoder[4] 구조를 선택했다. DeepCluster[3]는 self-supervised learning 중 pre-text task 단계의 특징 추출기를 잘 학습시키기 위해 특징 벡터에 군집화 알고리즘을 적용하여 반복적으로 입력 영상에 유사-레이블을 주는 방식으로 학습을 진행한다.

영상 군집화 연구의 성능을 표현할 때는 MNIST 손글씨 숫자 영상 데이터셋이 주로 사용된다[1,2]. 10개의 클래스에 대해 28×28 이라는 비교적 작은 차원으로 영상이 구성되어 군집화 성능을 평가하는데 비교적 용이하기 때문이다. 한편 고해상도 컬러 영상의 경우 차원의 크기가 기하급수적으로 커지기 때문에 입력 영상에서 저차원으로 압축된 특징 벡터를 잘 추출하는 것이 중요한 기준이 된다. 영상의 특징을 잘 압축하여 추출하지 못하면 좋은 군집화 성능을 기대하기는 어렵기 때문이다. 따라서 최근에는 고해상도 컬러 영상으로 이루어진 자연 영상에 대한 특징 추출 및 군집화 연구가 활발히 진행되고 있다.

본 논문에서는 자연 영상에 대한 Convolutional Auto Encoder의 특징 추출 성능을 평가하기 위해 설계한 실험 방법을 소개한다. 특히 모델의 특징 추출 능력을 순수하게 확인하기 위하여 Self-supervision 및 유사-레이블을 제공하지 않은 채 Naive한 모델의 결과를 분석한다. 실험을 위해 설계된 4가지 비지도학습 모델의 복원 결과를 통해 모델별 학습 정도를 확인한다. 그리고 비지도 모델이 다량의 unlabeled 영상으로 학습 되었음에도 불구하고 더 적은 labeled 데이터로 학습된 지도학습 모델의 특징 추출 성능에 못 미친다는 것을, 군집화 및 분류 실험 결과를 통해 살펴본다. 또한, 지도학습 모델에 데이터셋 간 교차 학습을 수행하여 모델에서 출력된 특징 벡터의 군집화 및 분류 성능도 확인한다. 각 실험에 사용된 특징 벡터들은 TSNE 시각화를 통해 모델별 특징 추출 양상을 확인할 수 있다. (부록 **그림2,3**)

2. 관련 연구

2.1. Convolutional Auto Encoder

Convolutional Auto Encoder(CAE)[4]는 데이터의 차원 압축 및 복원에 사용되는 Stacked Auto Encoder 구조에서, 완전연결계층 대신 합성곱 필터 계층을 사용하는 오토인코더 모델이다. 2D 영상에 대해 합성곱 신경망이 갖는 장점을 이용하였기 때문에, 높은 해상도의 영상에 대해서도 모델 파라미터가 잘 수렴한다는 특징이 있다. CAE의 구조는 영상의 크기를 줄이는 인코더(Encoder; downsampling path) 부분과 다시 크기를 키우는 디코더(Decoder; upsampling path) 부분이 순서대로 연결되어있다.

오토인코더는 구조적 특성상 인코더와 디코더 사이에서 잠재 벡터(latent vector) 형태를 거치게 된다. 이때 입력 데이터의 정보를 유지하며 차원이 감소된 잠재 벡터를 일종의 특징 벡터로 간주하고 사용할 수 있다. 특히 *E Blanco-Mallo et al.* [5]은 자신들의 추천 시스템 연구에서 CAE를 특징 추출기로 사용하는 것이 일반적인 합성곱 신경망을 통해 추출된 deep features를 사용하는 것보다 효과적이라고 주장한다.

한편 CAE의 압축-복원 과정에서 발생하는 정보손실로 인해 영상의 세밀한 부분들이 온전히 복원되지 못하기도 한다. 이와 같은 문제를 예방하기 위해 인코더에서 디코더로 동일 Scale의 특징 맵(Feature Map) 정보를 전달하는 skip architecture 방식이 추가될 수 있다. UNet[6] 모델은 해당 방식을 통해 Image Segmentation 분야에서 좋은 성능을 보여주었다.

2.2. ResNet

ResNet[7] 모델은 합성곱 신경망의 계층이 깊어짐에 따라 발생하는 gradient vanishing 문제를 해결하기 위해 고안된 분류 네트워크이다. 해당 모델은 특징 맵 사이의 skip connection을 이용한 잔차 학습(residual learning)을 바탕으로, 깊은 신경망 모델을 설계하는 방법론적 지평을 열었다. ResNet은 영상의 특징을 잡아내는데 우수한 성능을 갖추고 있어, 분류 문제 외에도 여타 문제들을 해결하는 모델의 백본(backbone) 네트워크로 자주 사용된다. 또한, 지도학습 데이터셋으로 학습된 ResNet 네트워크에서 Head 단을 제거하면 영상 특징 추출기로 사용할 수 있다.

3. 실험 설계

3.1. 실험 모델 구조

실험에 사용된 지도학습 모델은 흔히 합성곱 신경망의 백본(Backbone) 네트워크로 사용되는 ResNet[7]이다. ImageNet으로 사전학습된 네트워크를 각 데이터셋으로 전이학습 시킨 뒤 Head 단을 제거하여 특징 추출기로 사용하였다. 또한, 비지도학습 모델 4가지를 설계하였는데 (부록 **표3**), UNet[6] 기반 모델 및 Convolutional Auto Encoder[4] 기반 모델 3가지로 이루어져 있다. UNet 모델과 CAEv1 모델은 skip architecture의 유무로 구분되고, CAEv2는 인코더와 디코더 사이에 완전연결계층으로 구성된 bottleneck을 사용한다는 것이 특징적이다. 그리고 CAEv3는 *E Blanco-Mallo et al.*[5]이 제안한 논문에서 사용된 구조를 그대로 구현하였다. 지도학습 모델은 Cross-Entropy Loss를, 비지도학습 모델은 MSE Loss를 최소화하는 방향으로 학습되었다.

3.2. 군집화 및 분류 실험

모든 특징 추출기는 128차원의 특징 벡터를 출력하고, 추출된 특징 벡터는 이후 군집화 및 분류 성능 실험에 사용되었다. 특징 벡터 간의 거리 관계를 통해 특징 추출기의 성능을 평가하는 것을 목표로 하였기 때문에 특징 벡터 군집화에는 K-means 알고리즘을 사용되었고 분류에는 K Nearest Neighbor(KNN) 알고리즘이 사용되었다.

3.3. 실험 데이터

실험에는 두 가지 자연 영상 데이터셋(CIFAR10, STL10)이 사용되었다. 먼저 CIFAR10은 분류 지도학습에 사용되는 자연 영상 데이터셋으로, 10개 클래스에 대한 $32 \times 32 \times 3$ 크기의 영상으로 구성되어있다. 학습 및 테스트 샘플은 각각 50k, 10k 쌍으로 모두 레이블이 포함된 데이터이다. 그리고 STL10은 분류 지도학습 및 비지도학습에 사용되는 자연 영상 데이터셋으로, 10개 클래스에 대한 $96 \times 96 \times 3$ 크기의 영상으로 구성되어있다. 레이블이 포함된 지도학습 전용 학습 및 테스트 샘플은 각각 5k, 8k 쌍으로 구성되어있고, 비지도학습을 위한 unlabeled 영상 100k 개가 별도로 포함되어있다.

실험에 따라 지도학습 모델을 학습시킬 때는 CIFAR10 및 STL10의 지도학습 샘플을 사용했고, 비지도학습 모델을 학습시킬 때는 STL10의 비지도학습 전용 샘플을 사용했다. 또한 모든 실험에서 테스트는 레이블이 포함된 각 데이터셋의 테스트 샘플을 사용하여 진행되었다.

4. 실험 결과 및 분석

4.1. 비지도학습 모델 복원 결과

그림 1은 STL10 데이터셋 중 unlabeled 영상으로 학습된 비지도학습 모델의 복원 결과이다. 모델에 따라 원본(GT) 영상으로의 복원 정도가 다르지만, 복원을 잘하는 모델일수록 합성곱 파라미터를 잘 학습했다고 간주할 수 있다. 학습이 잘된 비지도학습 모델의 경우, 인코더 부분에서 입력 영상의 정보를 잘 유지하며 특징 벡터로 압축했다고 예상할 수 있다. **그림 1**에 따르면 UNet 및 CAEv1 모델이 원본 영상의 구조적 특성을 잘 복원한 것으로 보이고 따라서 특징 추출도 잘할 것으로 예상된다.

4.2. 모델별 특징 추출에 따른 군집화 및 분류 성능 결과

표 1은 지도학습/비지도학습 모델별 특징 추출에 따른 군집화 및 분류 성능 결과이다. 군집화 성능 평가 metric으로 Pairwise/Bicubed F-score는 Precision과 Recall의 조화평균을 계산하는 두 가지 방법을 나타낸다. 또한, Normalized Mutual Information (NMI)을 통해 군집

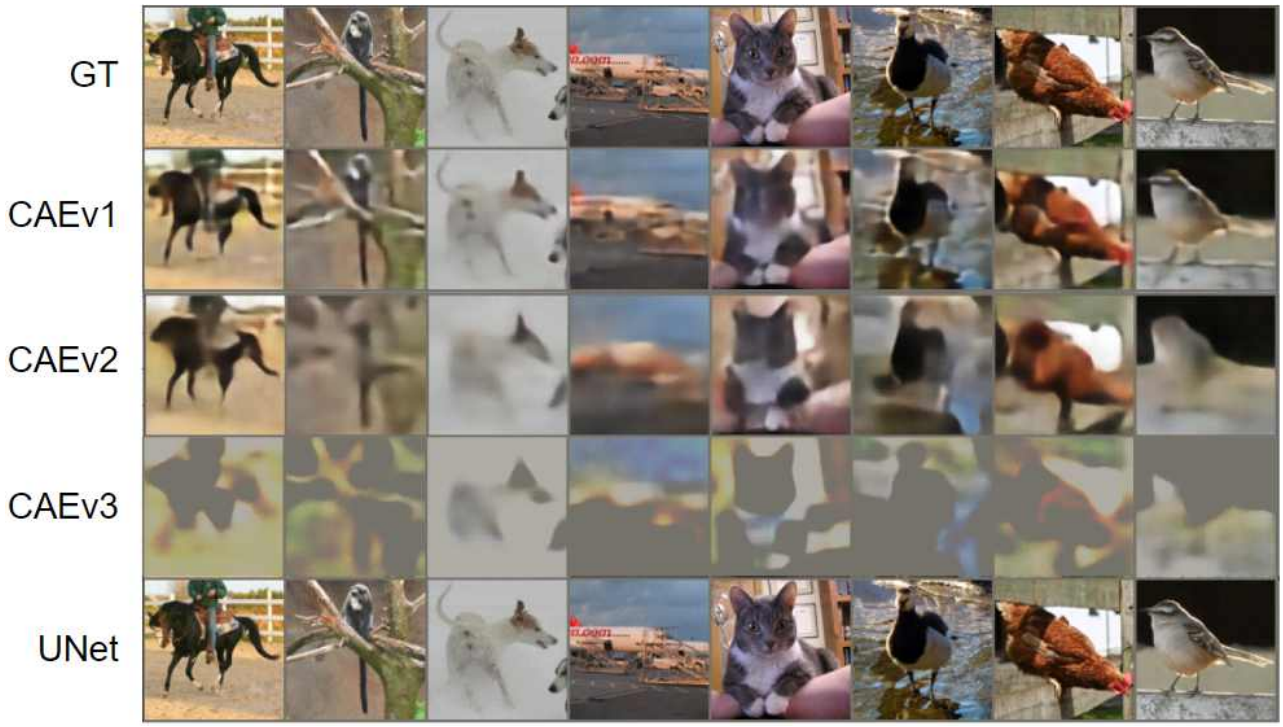

그림 1. 비지도학습 모델 복원 결과: STL10 unlabeled 영상으로 학습된 비지도학습 모델로 STL10 테스트 영상을 복원한 결과이다.

표 1. 모델별 특징 추출에 따른 군집화 및 분류 성능: 각 지도학습/비지도학습 모델로 추출한 특징 벡터를 이용하여 군집화 및 분류 실험을 수행한 결과이다. (동일 테스트 샘플: STL10 labeled test set, 8k)

		군집화 (K-means, K=10)			분류 (KNN, K=5)	비고
		Pairwise F-score	Bicubed F-score	NMI	Accuracy	
비지도학습 모델	CAEv1	0.1395	0.1481	0.0864	0.297	학습: STL10 unlabeled, 100k 테스트: STL10 labeled, 8k
	CAEv2	0.1715	0.1834	0.1538	0.313	
	CAEv3	0.1525	0.1598	0.1220	0.291	
	UNet	0.1827	0.1988	0.1341	0.296	
지도학습 모델	ResNet18	0.7986	0.7997	0.7847	0.892	학습: STL10 labeled, 5k 테스트: STL10 labeled, 8k
	ResNet34	0.8131	0.8131	0.7953	0.900	
	ResNet50	0.8329	0.8332	0.8163	0.911	

표 2. 교차 학습된 지도학습 모델의 군집화 및 분류 성능: 데이터셋을 교차하여 학습시킨 지도학습 모델의 군집화 및 분류 성능을 나타낸 표이다. (교차 학습: CIFAR10 labeled train set, 50k)

		군집화 (K-means, K=10)			분류 (KNN, K=5)	비고
		Pairwise F-score	Bicubed F-score	NMI	Accuracy	
비교차 학습	비지도 모델 최고 성능	0.1827	0.1988	0.1538	0.313	학습: STL10 unlabeled, 100k 테스트: STL10 labeled, 8k
	지도 모델 최저 성능	0.7986	0.7997	0.7847	0.892	학습: STL10 labeled, 5k 테스트: STL10 labeled, 8k
교차 학습	ResNet18	0.4536	0.4999	0.5239	0.694	교차 학습: CIFAR10 50k 테스트: STL10 labeled, 8k
	ResNet34	0.4408	0.4896	0.5138	0.700	
	ResNet50	0.4463	0.4869	0.5146	0.686	

화 평가에 사용되는 homogeneity와 completeness의 정도를 동시에 확인할 수 있다. 분류 성능 평가 metric으로는 원본 label과 얼마나 일치하는지를 나타낸 Accuracy를 사용하였다. 성능 평가에 사용된 모든 metric들은 0부터 1 사이의 값으로, 1에 가까울수록 좋은 성능을 의미한다.

결과에 따르면 비지도학습 모델이 더 많은 학습 영상으로 (STL10 unlabeled data, 100k 개) 학습했음에도 불구하고, 적은 학습 데이터로 (STL10 labeled, 5k 쌍) 학습한 지도학습 모델보다 군집화 및 분류 성능이 현저히 떨어지는 것을 확인할 수 있다. 이는 곧 비지도 모델의 인코더가 추출하는 특징 벡터의 정보 압축 품질이, 복원을 위한 학습을 얼마나 잘했는지와는 별개라는 것을 나타낸다. 해당 사실은 특징 벡터 시각화 자료(부록 **그림2**)를 통해 확인할 수 있는데, 특징 추출을 잘할 것으로 예상한 UNet 및 CAEv1 모델에서 특징 벡터가 클래스별로 모이지 않는 것을 볼 수 있다. 반면 지도 모델의 경우 특징 벡터가 비교적 잘 군집화되는 것을 확인할 수 있다.

추가로, 지도학습의 경우 ResNet 모델이 커질수록 군집화 및 분류 성능이 좋아지는 경향성을 확인할 수 있다. 이는 모델 파라미터가 많아질수록 합성곱 신경망의 표현력이 강해진다는 사실로부터 예상할 수 있는 결과이다.

4.3. 교차 학습된 지도학습 모델의 군집화 및 분류 성능 결과

표 2는 데이터셋을 교차하여 학습시킨 지도학습 모델의 군집화 및 분류 성능을 나타낸 표이다. 본 실험을 위해 학습에는 CIFAR10 train set (labeled, 50k 쌍)이 이용되었고, 테스트에는 **항목 4.2.**와 동일한 STL10 test set이 사용되었다. 결과에 따르면 교차 학습된 지도 모델의 성능은 **항목 4.2.**의 비지도 모델 최고 성능보다는 우수하고 지도 모델 최저 성능보다는 좋지 않다는 것을 나타낸다. 해당 내용은 교차 학습 모델의 특징 벡터 시각화 자료(부록 **그림3**)를 통해 보충 설명될 수 있다. 이는 곧 비지도 모델이 자연 영상에서 추출하는 특징 벡터의 품질이 예상보다 많이 안 좋다는 것을 의미한다. 즉, 목표로 하는 데이터 Domain의 unlabeled 자연 영상으로 비지도 모델을 학습하여 사용하는 것보다, 다른 자연 영상 Domain으로 사전 학습된 지도 모델을 사용하는 것이 특징 벡터를 추출하여 군집화 및 분류를 하는 것에 유리함을 알 수 있다.

또한, 교차 학습된 ResNet의 경우 모델의 크기와 특징 벡터의 군집화 및 분류 성능 간에는 특별한 상관관계가 없는 것으로 보인다. 이는 학습/테스트 데이터셋 사이의 Domain Gap이 모델의 표현력으로 커버할 수 있는 범위를 넘었기 때문이라고 해석할 수 있다.

5. 결론

본 논문에서는 자연 영상에 대한 Naive Convolutional Auto Encoder의 특징 추출 성능을 평가하기 위해 3가지 실험을 설계했다. 먼저 4가지 비지도학습 모델을 다량의 unlabeled 자연 영상으로 학습시킨 뒤 원본 복원 결과를 확인했다. 또한, 적은 양의 labeled 데이터로 학습된 지도학습 모델과의 특징 추출 능력을 비교하기 위해 특징 벡터의 군집화 및 분류 실험을 진행했다. 마지막으로 데이터셋을 교차 학습시켰을 때 지도학습 모델의 특징 추출 능력을 확인하기 위해 동일 실험을 진행하였다.

실험 결과에 따라 비지도 모델이 복원을 잘하는지와 인코더가 특징 벡터를 잘 압축하여 추출하는지는 크게 상관이 없음을 알 수 있었다. 특히 더 적은 데이터로 학습한 지도 모델의 특징 벡터 추출 능력이 비지도 모델의 추출 능력보다 뛰어나다는 사실을 군집화 및 분류 성능 결과를 통해 확인할 수 있었다. 또한, 비지도 모델을 사용하는 것보다 차라리 다른 자연 영상 Domain으로 교차 학습된 지도 모델을 사용하는 것이 군집화 및 분류를 위한 특징 벡터를 추출하는데 더 유리하다는 것을 알 수 있었다.

감사의 글

본 논문은 중소벤처기업부의 2020년 AI기반 고 부가 신제품기술개발 사업의 지원을 받아 수행함(주관기관: ㈜자비스). 이 연구는 2021년 산업통상자원부 및 산업기술평가관리원(KEIT) ATC+ 사업의 연구비 지원에 의한 연구임(과제번호: 20014131, 반도체 후공정불량검사를 위한 AI 기반 25mm급 X-ray 자동 검사 시스템 개발). 이 논문은 2022년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. “This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022."

Reference

[1] J Xie et al. “Unsupervised Deep Embedding for Clustering Analysis.” In Proceedings of the 33rd International Conference on Machine Learning (ICML), PMLR 48:478-487, 2016.

[2] X Guo et al. “Deep clustering with convolutional autoencoders.” In International Conference on Neural Information Processing (ICONIP), pp 373-382, 2017.

[3] M Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features.” In Proceedings of the European Conference on Computer Vision (ECCV), pp 132-149, 2018.

[4] J Masci et al. “Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction.”, In International Conference on Artificaial Neural Networks (ICANN), pp 52-59, 2011.

[5] E Blanco-Mallo et al. “On the effectiveness of convolutional autoencoders on image-based personalized recommender systems.” In Proceedings of 3rd XoveTIC Conference, 2020.

[6] O Ronneberger et al. “U-net: Convolutional networks for biomedical image segmentation.” In Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp 234-241, 2015.

[7] K He et al. “Deep Residual Learning for Image Recognition.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770-778, 2016.