데이터셋 응축 연구

요 약

데이터셋 응축은 거대해지는 이미지 데이터셋의 효율을 높이기 위해 학습 정확도는 유사하지만 크기는 아주 작은 데이터셋을 합성하는 연구 분야이다. 본 연구에서는 제안된 여러 데이터셋 응축을 면밀히 분석하고 기존 성능 향상이 미미했던 의미 분석(semantic analysis)을 기반한 데이터 증강 기법을 개선한다. 나아가 제안된 여러 데이터셋 응축 방법들은 원본 데이터 공간과 유사한 응축 샘플이 높은 학습 정확도를 갖는 것을 보였다. 본 연구 또한 원본 데이터 공간과 유사하지만 최대한 다양한 응축 샘플들을 만들기 위한 방법을 모색한다. 응축 데이터셋은 데이터 효율 뿐만 아니라 사회, 환경 관점에서 긍정적인 기여를 할 수 있기에 원본 데이터 보호, 소모 에너지양을 분석하여 응축 데이터셋의 사회적 측면 또한 분석한다.

1. 서론

1.1. 연구배경

Deep Neural Network 를 시작으로 여러 기계학습 모델은 이미지 처리, 음성 인식 등 여러 분야에서 최고의 성능을 보여주며 이들이 사용되는 분야 또한 확대되고 있다. 데이터셋은 기계학습을 다양한 분야에 적용할 수 있도록 뒷받침하며 그 질과 양이 기계학습 기술과 함께 발전하고 있다. 특히, 데이터의 질과 양이 모델 성능에 큰 영향을 미치기 때문에 높은 성능을 내기 위해 방대한 데이터셋을 여러 분야에서 활용하고 있다[1]. 용량이 큰 데이터셋은 저장하는 것도 어려울 뿐만 아니라 이를 활용하기 위해 높은 컴퓨팅 성능을 요구한다. 결국 제한된 컴퓨팅 자원으로 사용할 수 있는 데이터의 종류와 양은 한계가 존재하고 이는 평생학습(Continual Learning)[2], Neural Architecture Search(NAS)[3]와 같이 다양한 데이터가 요구되는 분야에서 어려움으로 나타난다.

데이터셋 응축(Dataset Condensation)은 제한된 저장 용량과 컴퓨팅 자원으로 학습 데이터의 효율을 극대화하기 위해 원본 데이터셋의 학습 정확도와 유사하도록 아주 작은 데이터셋(e.g. 10 images/class)을 생성하는 분야이다. Zhao et al(2018b)[1]에서는 응축한 데이터가 실제 데이터와 유사한 가중치를 발생시키도록 합성하는 방법이 제안되었고 이를 토대로 다양한 데이터셋 응축 연구가 활발히 진행되고 있다. 특히, 입력 이미지에 attention map 을 사용해 마스킹하여 샘플을 합성하는 방법은 핵심 정보에 집중하여 데이터셋 응축을 수행한다[3]. 그러나 마스킹하는

데이터가 입력 이미지와 연관성이 적어 핵심 정보에 집중했다는 설득력이 부족하다. 본 연구에서는 기존 제안된 attention map 을 이용한 데이터셋 응축을 개선하고자 한다.

1.2. 연구목표

본 연구는 데이터셋 응축 성능을 극대화하기 위해 데이터와 모델 측면에서 여러 방법을 살펴보고 데이터셋 응축에 핵심적인 역할을 하는 요소를 분석한다. 이를 바탕으로 기존 제안된 attention map 을 이용한 데이터셋 응축의 문제점을 면밀히 분석한다. 나아가 문제점을 해결하기 위해 데이터 증강, 대조 학습 등의 연구 분야에서 사용되는 방법론을 적용한다.

데이터셋 응축은 그 가치가 데이터 효율성 증가에도 있지만 원본 데이터 보호, 학습 시 사용되는 에너지 절약 등 다양한 사회적 가치를 갖는다. 본 연구를 통해 응축한 데이터셋이 갖는 사회적 가치를 분석하여 응축 데이터셋의 효용을 살펴본다.

2. 관련연구

2.1. 데이터셋 응축

Wang et al. (2018)[5]은 처음으로 학습 가능한 데이터셋 응축 제안하며 이을 이중 수준 최적화(Bi-Level Optimization) 문제로 설계하여 유의미한 결과를 얻었다. Zhao et al. (2021)[1]는 이전 연구를 확장해 응축된 이미지 샘플들이 원본 데이터와 유사한 학습 과정을 만들어내도록 합성하는 방법을 제안했다. 이를 위해 layer 에서 발생하는 가중치를 사용하는 gradient matching loss 를 제안했다. Zhao & Bilen(2021)[6]은 같은 학습 방법론에 미분 가능한 데이터 증강을 적용하여 데이터 효율성을 극대화하는 방법을 제안했다.

한편 응축 데이터셋이 원본 데이터셋과 유사한 데이터 공간에 위치하도록 학습하는 데이터셋 응축 방법이 제안되었다[7]. Gradient matching loss 를 사용한 데이터셋 응축보다 성능이 낮은 경우(i.e. MNIST)가 있지만 연산이 줄어 ImageNet 까지 확장할 수 있었다. 이와 유사하게 원본 데이터 분포 내에서 최대한 다양한 샘플을 생성하도록 discriminative loss 를 사용한 데이터셋 응축 방법이 제안되었다[8]. Kim et al. (2022)[9]는 하나의 응축 샘플에 데이터 증강을 적용해 여러개의 응축 샘플을 생성하고 이를 사용하여 데이터셋 응축을 수행한다. Lee et al (2022)[10]은 이와 유사하게 학습 가능한 은닉 공간에서의 *code* 와 간단한 *decoder* 를 사용해 더 다양한 응축 이미지를 생성해낸다. 이외에도 Infinite Wide Neural Network 와 Kernel Ridge Regression 을 이용해 데이터셋을 응축하지만 수백대의 GPU 를 요구한다는 한계점이 있다[11].

본 연구는 위 연구들을 분석해 사용한 loss, 데이터 증강 기법에 따른 응축 데이터셋의 데이터 공간을 살펴보고 정확도 향상에 대한 상관관계를 분석한다. 이를 바탕으로 기존 attention map 을 이용한 데이터셋 응축 성능 향상에 적용한다.

3. 프로젝트 내용

3.1. 데이터 측면

데이터셋 응축에서 데이터 증강의 효과는 Zhao & Bilen(2021)[6]의 연구 등에서 그 효과를 입증했다. 본 연구에서는 기존 제안되었던 attention map 을 이용한 마스킹과 배경정보를 제거한 데이터로 데이터셋 응축을 수행하고 응축 데이터셋의 데이터 공간 분포를 확인한다. 나아가 Zhao & Bilen(2021)의 방법의 결과와 비교하여 attention map 을 이용한 마스킹과 배경정보 제거의 효용을 분석한다.

학습 데이터를 의미 분석(semantic analysis)을 적용해 데이터셋을 응축한 시도는 아직 이루어지지 않았다. 따라서 위 연구에서 분석한 내용을 바탕으로 이미지 데이터 특징 학습을 극대화 할 수 있는 의미 분석에 기반한 데이터 증강 방안을 모색한다.

3.2. 모델 및 학습 측면

다양하고 효율 높은 응축 데이터셋을 만들기 위해 matching loss[1], discriminative loss[8] 등이 제안되었다. 특히 같은 부류의 데이터에 대해 원본 데이터의 분포에 최대한 골고루 응축 데이터셋을 합성하도록 제한하는 discriminative loss 는 다른 연구와 결합해 이의 효용을 살펴보기 좋다. 나아가 부류 간 은닉 데이터 공간을 넓히는 대조 학습(contrastive learning)[12]의 적용 가능성을 알아본다.

4. 향후 일정

10 월 동안 기존 제안되었던 데이터셋 응축 방법들을 검증한다. 또한 기존 attention map 을 이용해 학습 데이터를 마스킹하고 데이터셋 응축하는 방법을 보완할 수 있는 학습 방법을 적용해보며 그 효과를 측정한다. 이를 통해 데이터셋 응축 성능 향상에 있어 핵심적인 요소를 발견하고 강화할 수 있는 방법에 대해 연구한다.

11 월 동안 새로 제안하는 방법에 대해 조사 및 실험을 수행한다. 위 실험은 CIFAR10 데이터셋을 비롯한 다양한 이미지 데이터셋을 사용한다. 또한 데이터셋의 적용 가능성을 살피기 위해 여러 모델에 교차 검증을 수행하며 제안한 방법을 면밀히 조사한다.

12 월 동안 수행했던 실험을 정리하고 새롭게 분석한 내용을 바탕으로 논문을 작성한다. 또한 개인정보 보호, 에너지 효율과 같이 성능 뿐만 아니라 사회와 환경 측면에서 제안한 방법을 검토하여 데이터셋 응축의 효용을 살핀다.

5. 결론 및 기대효과

본 연구를 통해 데이터 증강 기법과 모델의 학습 능력에 따른 데이터셋 응축 성능을 비교한다. 다양한 실험을 통해 데이터셋 응축 성능 향상에 핵심적인 역할을 수행하는 요소를 분석하고 기존 제안되었던 방법을 보완하기 위한 연구를 수행한다. 나아가 실험 결과를 성능 측면 뿐만 아니라 사회 및 환경 측면에서 분석하여 이의 효용을 면밀히 검토한다. 마지막으로, 제안한 데이터셋 응축 기법으로 생성된 데이터셋을 배포하여 다양한 분야에 사용할 수 있도록 한다.

본 연구에서 제안하고자 하는 효과적이고 강력한 데이터셋 응축 방법은 평생 학습(Continual Learning), Neural Architecture Search(NAS)와 같은 분야의 발전을 가속화할 수 있다. 또한 원본 데이터셋의 보안을 유지하면서 데이터의 활용을 극대화 할 수 있고 적은 에너지로 충분히 학습이 가능하다. 본 연구가 사회의 안전, 에너지 절약, 그리고 여러 분야의 발전에 도움을 줄 것으로 보인다.

6. 참고문헌

[1] Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. "Dataset condensation with gradient matching." arXiv preprint arXiv:2006.05929 (2020).

[2] Liu, Hanxiao, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." arXiv preprint arXiv:1806.09055 (2018).

[3] Lopez-Paz, David, and Marc'Aurelio Ranzato. "Gradient episodic memory for continual learning." Advances in neural information processing systems 30 (2017).

[4] Donghoon Kim, Sungho Bae, “Dataset Condensation using Attention”, 34th Workshop on Image Processing and Image Understanding(IPIU), 2022

[5] Wang, Tongzhou, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. "Dataset distillation." *arXiv preprint arXiv:1811.10959*(2018).

[6] Zhao, Bo, and Hakan Bilen. "Dataset condensation with differentiable siamese augmentation." In *International Conference on Machine Learning*, pp. 12674-12685. PMLR, 2021.

[7] Zhao, Bo, and Hakan Bilen. “Dataset Condensation with Distribution Matching.” *arXiv preprint arXiv:2110.04181(2022)*

[8] Wang, Kai, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. "Cafe: Learning to condense dataset by aligning features." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12196-12205. 2022.

[9] Kim, Jang-Hyun, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. "Dataset Condensation via Efficient Synthetic-Data Parameterization." *arXiv preprint arXiv:2205.14959* (2022).

[10] Lee, Hae Beom, Dong Bok Lee, and Sung Ju Hwang. "Dataset Condensation with Latent Space Knowledge Factorization and Sharing." *arXiv preprint arXiv:2208.10494*(2022).

[11] Nguyen, Timothy, Roman Novak, Lechao Xiao, and Jaehoon Lee. "Dataset distillation with infinitely wide convolutional networks." *Advances in Neural Information Processing Systems* 34 (2021): 5186-5198.

[12] Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In *International conference on machine learning*, pp. 1597-1607. PMLR, 2020.