인공지능 모델의 학습 데이터 영향 분석: 영향 함수의 활용과 한계에 관한 연구

고정연, 양현종*

포항공과대학교

jwe06020@postech.ac.kr, *hyunyang@postech.ac.kr

A Study on Analyzing the Influence of Training Data on AI Models: Utilization and Limitations of Influence Functions

Jungyeon Koh, Hyunjong Yang*

Pohang University of Science and Technology (POSTECH)

요 약

학습 데이터가 인공지능 모델에 미치는 영향을 어떻게 설명할 수 있을까? 영향 함수 (Influence Function)는 모델의 그라디언트와 헤시안 행렬만을 이용하여 이에 대한 해결책을 제공한다. 그러나 영향 함수는 여전히 대규모 및 비선형 모델에 적용할 경우 메모리와 계산 비용이 급증하고, 추정 정확도가 떨어지는 단점이 있다. 본 논문은 영향 함수의 개념과 이러한 한계를 극복하기 위한 관련 최신 연구 동향을 분석해보고자 한다.

Ⅰ. 서론

지난 10여 년 동안 신경망 구조의 놀라운 발전 속에서 인공지능 알고리즘은 눈부시게 성장하였다. 더 깊은 레이어와 많은 수의 매개변수를 갖춘 GPT 시리즈는 현재 약 1조 개의 파라미터를 자랑하며, 인공지능 기술이 곧 인간 지성의 영역에 도달할 수 있는 잠재력을 갖췄음을 시사한다. 인공지능 모델의 규모가 커지면서 그 활용성은 확장되었지만, 반면에 모델의 설명력을 분석하는 것은 제한된 컴퓨팅 자원으로 인해 더욱 어려워졌다. 따라서 차세대 초대형 인공지능의 효율적 개발 및 운용을 위해서는 훈련된 신경망의 동작을 이해하고 오류를 탐지하여 처리하는 방법이 필요하다.

영향 함수는 고전 통계학에서 유래되었으며, 최근 논문 [1]을 통해 인공지능 모델의 설명력을 제공하는 중요한 기술적 발전으로 부상했다. 이 방법은 모델의 그라디언트와 헤시안 행렬만을 기반으로 특정 훈련 데이터를 제거할 때 모델 추론 능력에 미치는 영향을 추정한다. 영향 함수를 사용하면 leave-one-out(LOO) 재학습 없이도 모델 성능에 관한 학습 데이터의 영향을 간단하게 평가할 수 있다는 장점이 있다. 이에 본 논문에서는 영향 함수의 개념 및 연구 동향을 분석하고 해결해야 할 문제들을 제안한다.

Ⅱ. 본론

A. 영향 함수의 개념

매개변수 공간 $\Theta \in \mathbb{R}^p$ 내 정의된 모델 θ와 n개의 훈련 데이터 $z_1, \ldots, z_n$가 주어졌을 때, 경험적 위험은 $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(z_i;\theta)$로 정의된다. 이때 영향 함수는 특정 데이터 z가 충분히 작은 ϵ만큼 가중되었을 때의 매개변수의 변화를 추적한다. ϵ-가중된 경험적 위험 최소화자는 아래와 같다.

$$\hat{\theta}_{\epsilon,z} = \arg\min_{\theta\in\Theta} L(\theta) + \epsilon l(z;\theta)$$

이후 1차 테일러 근사 및 뉴턴-랩슨법을 적용하면 아래와 같은 영향 함수를 구할 수 있다.

$$I(z) := \left.\frac{d(\widehat{\theta_{\epsilon,z}} - \hat{\theta})}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta l(z;\theta)$$

영향 함수를 위와 같이 계산하려면 헤시안의 역행렬이 필요하며, 이는 시간 복잡도가 $O(p^3)$인 엄청난 계산 병목을 발생시킨다. 이에 따라 논문 [1]은 헤시안의 역행렬 대신 inverse-Hessian-vector product (iHVP)를 사용하는 Linear-time Stochastic Second-order Algorithm (LiSSA) 반복 알고리즘을 제안한다.

$$I_k = I_0 + (\mathbb{I} - H_{\hat{\theta}})I_{k-1}$$

이때 $\mathbb{I} \in \mathbb{R}^{p\times p}$는 항등행렬이며, $H_{\hat{\theta}} := \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta{}^2 l(z_i;\hat{\theta}) \in \mathbb{R}^{p\times p}$, $I_0 = \nabla l(z;\hat{\theta})$이다. 반복 알고리즘은 미리 정의된 임계값 δ에 대해 $\|I_k - I_{k-1}\| \leq \delta$를 만족할 시 종료된다.

B. 영향 함수의 최신 연구 동향 및 한계점 분석

영향 함수는 인공지능 모델 추론 능력을 해석함으로써 설명 가능한 인공지능의 새로운 지평을 열었지만, 여전히 두 가지의 치명적인 단점을 가지고 있다. 첫째는 상당한 메모리 및 계산 오버헤드이며, 두 번째는 대규모 모델 및 데이터셋에 적용할 때의 부정확성이다. 대규모 언어 모델 (LLM)과 같은 초대형 인공지능 시스템이 등장함에 따라 이러한 영향 함수의 단점은 응용 분야를 확장하는 데에 있어 큰 장애물이 되고 있다.

LiSSA 알고리즘은 여전히 $O(nrp)$의 시간 복잡도를 갖는다. 이때 n은 분석할 훈련 데이터의 수, r은 필요한 에폭 횟수이며, $p = |\Theta|$이다. 논문 [2]는 매개변수의 일부만을 이용하여 $O(p)$의 시간 복잡도를, 논문 [3]은 데이터셋의 일부만을 이용하여 $O(n)$의 시간 복잡도를 완화하였다. 그러나 이러한 접근 방식은 알고리즘의 선택적 성격 때문에 예상치 못한 편향을 초래할 수 있다.

또한 영향 함수는 대규모 데이터셋이나 심층 신경망에 적용 시 LOO 재학습의 영향을 부정확하게 평가하는 경향이 있다. 영향 함수는 신경망의 손실 함수가

strongly convex하다고 가정한다. 이에 따라 선형 회귀 모델, Soft Vector Machine (SVM)과 같은 단순한 모델에서는 좋은 성능을 보이지만, 심층 신경망을 사용하는 최신 인공지능 모델에서는 그만큼 뛰어난 성능을 내지 못하고 있다. 한편, 논문 [4]는 영향 함수의 내재된 변동성을 다섯 개의 요인으로 분류하여 분석하였다. 그 중 가장 큰 비중을 차지하는 것은 warm-start gap이며, 이는 매개변수 최적점 (θ^*)에 위치한 warm-start 상태의 모델이 그 최적점에 편향되어 있음을 의미한다. 이에 따라 최신 심층 신경망의 경우 최적점의 비특이성으로 인해 영향 함수가 보다 큰 변동성을 갖는다.

이를 해결하기 위해 논문 [1]은 iHVP에 아래와 같이 감쇠항 λ를 도입하여 고유값을 양수로 유지하였다.

$$\tilde{L}(z,\theta) = L(z,\tilde{\theta}) + \nabla L(z,\tilde{\theta})^\top(\theta-\tilde{\theta})^\top + \frac{1}{2}(\theta-\tilde{\theta})(H_{\tilde{\theta}}+\lambda I)(\theta-\tilde{\theta})$$

감쇠항의 도입은 L_2 정규화와 유사한 효과를 나타내며, 손실 함수의 볼록성을 유지하는데 도움을 준다. 하지만 실제 L_2 정규화를 적용하기 전에 적절한 λ를 결정하는 것이 어렵기 때문에 영향 함수의 변동성을 근본적으로 해결하지는 못한다. 논문 [5]에서는 영향 함수의 정확도를 개선하기 위해 2차 테일러 근사를 적용했지만, 계산 복잡도가 급격히 증가함에 따라 실제 모델에 대한 적용 가능성이 제한된다. 마찬가지로, 논문 [6]에서는 헤시안 행렬을 보다 정확하게 근사하기 위해 피셔 정보 행렬을 사용했지만, 이 또한 상당한 계산 오버헤드를 초래한다.

C. 영향 함수의 응용 분야

영향 함수는 다양한 분야에 활용될 수 있다. 논문 [2]는 영향 함수를 모델 디버깅, 학습 데이터셋 품질 관리 및 적대적 학습 데이터 생성에 사용하였다. 최근 연구들은 더 나아가 영향 함수를 학습 데이터 증강 [7], 데이터셋 프루닝[8]에도 적용하였다.

기존 연구들이 주로 약 10억 개의 매개 변수를 갖는 모델에 영향 함수를 적용한 반면, 논문 [9]는 EK-FAC (Eigenvalue-Corrected Kronecker-Factored Approximate Curvature)와 TF-IDF (Term Frequency-Inverse Document Frequency)을 사용하여 영향 함수를 520억 개의 매개변수를 갖는 대규모 언어 모델에 적용하였다. 이는 영향 함수가 줄곧 한계점으로 여겨졌던 초대형 인공지능 모델에서의 제한성을 극복했음을 시사한다. 또한 이 논문은 영향 함수를 이용하여 프롬프트 p와 답변 c에 대하여 $\log \Pr(c|p)$에 가장 큰 영향을 끼친 학습 데이터 시퀀스를 $O(mp)$의 시간 복잡도 내에 탐색하였다. 이때 m은 학습 데이터의 차원을 뜻한다.

이처럼 영향 함수는 인공지능 모델 및 데이터셋 디버깅에 도움을 제공한다. 영향 함수의 계산 및 메모리 오버헤드를 개선하기 위한 알고리즘의 개발은 보다 큰 신경망에서의 영향 함수의 응용 가능성을 확장하는데 큰 기여를 할 것이다.

Ⅲ. 결론

초대형 인공지능 모델은 우리가 상상할 수 없는 능력을 보여주고 있다. 하지만 여전히 우리는 그 동작을 이해하지 못해 성차별, 인종차별 등의 추론 편향이나 적대적 공격에 대처하는 데 어려움을 겪는다. 모델의 행동을 설명하는 데 있어 훈련 데이터의 시각을 이용하는 영향 함수는 인공지능 모델의 기계적 이해를 위한 상향식 접근법의 초석으로 해석할 수 있다. 모델의 세부 회로 및 기능을 이해하는 하향식 방식과 영향 함수를 기반으로 한 상향식 방식을 모두 활용한다면, 우리는 완전한 이해를 기반으로 보다 견고한 인공지능 모델을 구축할 수 있을 것이다.

ACKNOWLEDGMENT

This research was supported in part by the IITP(Institute for Information & Communications Technology Planning & Evaluation), grant funded by MSIT(Ministry of Science and ICT) (RS-2024-00229541), in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00250191), and in part by the MSIT, Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2021-0-02048) supervised by the IITP.

참 고 문 헌

[1] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In International conference on machine learning, pages 1885-1894. PMLR, 2017.

[2] Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, and Pradeep Ravikumar. First is better than last for language data influence. Advances in Neural Information Processing Systems, 35:32285-32298, 2022.

[3] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. arXiv preprint arXiv:2012.15781, 2020.

[4] Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B Grosse. If influence functions are the answer, then wjhat is the question? Advances in Neural Information Processing Systems, 35:17953-17967, 2022.

[5] Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In International Conference on Machine Learning, pages 715-724. PMLR, 2020.

[6] Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. Interactive label cleaning with example-based explanations. Advances in Neural Informatio Processing Systems, 34:12966-12977, 2021.

[7] Donghoon Lee, Hyunsin Park, Trung Pham, and Chang D Yoo. Learning augmentation network via influence functions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10961-10970, 2020.

[8] Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. arXiv preprint arXiv:2205.09329, 2022.

[9] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large lanugage model generalization with influence functions. arXiv preprint arXiv: 2308.03296, 2023.