과최적화로 인한 표현 붕괴의 분야 간 공통 양상에 대한 고찰

이승민, 이정우*
서울대학교

seungmin7792@cml.snu.ac.kr, junglee@snu.ac.kr

A Study on Cross-Domain Common Patterns of Representation Collapse Induced by Over-Optimization

Lee Seung min, Lee Jung Woo*
Seoul National Univ., Seoul National Univ.

요 약

본 논문은 강화학습, 인간 피드백 기반 강화학습, 대규모 언어모델 학습 전반에서 과최적화로 인해 내부 표현이 붕괴되는 현상을 공통 문제로 정리하고, 분야별로 상이한 용어와 지표에 의해 분절적으로 기술·계측되는 현황이 통합적 이해의 정립을 어렵게 만든다는 문제의식을 제기한다. 이를 위해 본 논문은 분야에 독립적으로 적용 가능한 과최적화 유발 표현 붕괴의 정량적 정의를 제안하며, 분야 간 비교 가능한 논의를 가능하게 하는 공통 논의 기반을 제시한다.

Ⅰ. 서 론

최근 강화학습, 인간 피드백 기반 강화학습, 대규모 언어모델 학습은 각기 다른 목적 함수와 학습 패러다임을 기반으로 급속히 발전해왔다. 한편 이들 세 연구 분야에서 공통적으로, 학습이 특정 목표를 강하게 최적화하는 과정에서 모델 내부 표현이 붕괴하는 양상이 보고되고 있다.

이러한 표현 붕괴는 분야별로 서로 다른 형태와 용어로 기술된다. 예컨대 강화학습에서는 가치함수 표현력의 저하 [1], 인간 피드백 기반 강화학습에서는 보상 신호에 대한 과도한 적응 [2], 대규모 언어모델 학습에서는 비등방성 증가 [3]와 같은 현상이 보고되어 왔다. 이렇듯 유사한 현상임에도 서로 다른 용어와 측정 지표를 통해 분절적으로 기술되고 있어, "표현 붕괴가 무엇이며 어느 정도 발생했는가"를 분야 간에 일관되게 논의하기 어렵다.

본 논문은 이러한 문제의식에서 출발하여, 강화학습, 인간 피드백 기반 강화학습, 대규모 언어모델 학습 전반에서 과최적화로 인해 내부 표현이 붕괴되는 현상을 공통 문제로 정리하고 이를 "과최적화로 인한 표현 붕괴"로 개념화한다. 나아가 각 분야에 독립적으로 적용 가능한 표현 붕괴의 정량적 정의를 제안함으로써, 서로 다른 분야에서 보고된 현상을 동일한 언어와 척도 위에서 논의할 수 있는 공통 논의 기반을 제시한다.

Ⅱ. 본론

본 논문이 다루는 "과최적화로 인한 표현 붕괴"는, 학습이 특정 목표를 강하게 최적화하는 과정에서 모델 내부 표현이 저차원화, 유사도 증가, 특징 소실 등의 형태로 퇴화하는 현상을 포괄한다. 본 논문에서 제시하는 "과최적화로 인한 표현 붕괴"의 정량적 정의는 아래와 같다.

$$O_t = (P_t, \varphi_{\theta_t}, D_{probe}, C_t)$$

O_t 는 학습 시점 t 에서 과최적화로 인한 표현 붕괴 상태를 정량적으로 기술하기 위한 관측 벡터이며, P_t 는 시점 t 에서 모델이 학습을 통해 최적화하고 있는 목적의 진행도를 나타낸다. φ_{θ_t} 는 시점 t 의 모델 파라미터 θ_t 하에서 내부 표현 벡터 z 를 추출하는 함수이다, $z = \varphi_{\theta_t}(x), x \in D_{probe}$. D_{probe} 는 표현 붕괴를 측정하기 위해 사용하는 고정된 입력 데이터 집합이다. C_t는 시점 t 에서 D_{probe} 에 대해 추출된 표현 집합, $Z_t = \{\varphi_{\theta_t}(x) | x \in D_{probe}\}$으로부터 계산되는 표현 붕괴의 정량 지표이다.

표현 붕괴의 정도를 정량적으로 나타내는 지표인 C_t는 $C_{dim}(t), C_{sim}(t), C_{inact}(t)$ 세가지 요소로 구성된다. 각 요소가 뜻하는 바는 다음과 같다. $C_{dim}(t)$는 시점 t 에서 D_{probe}에 대한 내부 표현들을 모았을 때, 그 표현 분포가 얼마나 저차원 부분공간으로 압축되어 있는지를 나타내는 지표이다. $C_{sim}(t)$ 는 시점 t 에서 서로 다른 입력들에 대응하는 내부 표현들이 서로 얼마나 유사한지를 나타내는 지표이다. $C_{inact}(t)$는 시점 t 에서 내부 표현의 각 특징 차원 중 입력에 따라 거의 변하지 않아 사실상 비활성화된 차원의 비율을 나타내는 지표이다.

위와 같이 정의된 O_t 는 강화학습, 인간 피드백 기반 강화학습, 대규모 언어모델 학습 모두에서 정의 가능하다. 정의된 O_t 와 실제 달성하려는 목표인 G_t 를 활용하여,

"과최적화로 인한 표현 붕괴" 현상이 일어났는지 아래와 같이 진단할 수 있다.

1) 약한 표현 붕괴 구간

$$\Delta P_t \geq \varepsilon_P \,, \Delta C_t \geq \varepsilon_C \quad \forall\, t \in [t_1, t_2]$$
$$t_2 - t_1 \geq L, \quad \varepsilon_P, \varepsilon_C > 0$$

2) 강한 표현 붕괴 구간

$$\Delta P_t \geq \varepsilon_P \,, \Delta C_t \geq \varepsilon_C, \Delta G_t \leq \varepsilon_G \quad \forall\, t \in [t_1, t_2]$$
$$t_2 - t_1 \geq L, \qquad \varepsilon_P, \varepsilon_C > 0, \varepsilon_G \leq 0$$

길이가 최소 L 이상인 구간 동안 최적화 목적의 진행도가 증가하며 최적화가 정상적으로 일어나고 있음에도 불구하고 표현 붕괴 정도가 증가할 경우에, 약한 "과최적화로 인한 표현 붕괴"가 발생했다고 판단한다. 이와 동시에, 강화학습에서 평가 리턴과 같은 실제 달성하고자 하는 목표인 G_t 에서 멀어질 경우 강한 "과최적화로 인한 표현 붕괴"가 발생했다고 판단한다.

Ⅲ. 결론

본논문에서는 강화학습, 인간 피드백 기반 강화학습, 대규모 언어모델 학습 전반에서 과최적화로 인해 내부 표현이 붕괴되는 현상을 공통 문제로 정리하고, 분야별로 상이한 용어와 지표로 분절적으로 기술되어 온 현황이 통합적 이해의 정립을 어렵게 만든다는 문제의식을 제기하였다. 이에 대한 대응으로 본 논문은 과최적화로 인한 표현 붕괴를 분야에 독립적으로 기술할 수 있는 정량적 정의를 제안한다. 더 나아가, 본 논문은 약한 표현 붕괴 구간과 강한 표현 붕괴 구간을 구분하는 진단 기준을 제시함으로써 과최적화로 인한 표현 붕괴가 목표 함수 최적화가 곧 성능 향상으로 이어지진 않는다는 괴리와 결합될 수 있음을 명시적으로 다루었다.

본 논문의 한계와 향후 과제는 다음과 같다. 첫째, D_{probe} 의 구성과 φ_{θ_t} 의 선택에 따라 붕괴 지표의 민감도와 해석이 달라질 수 있으므로, 표준화된 원칙이 추가로 정교화될 필요가 있다. 둘째, $C_{dim}(t), C_{sim}(t), C_{inact}(t)$ 의 구체적 계산 방식 및 임계값($\varepsilon_P, \varepsilon_C, \varepsilon_G, L$) 설정은 연구 목적과 환경에 따라 달라질 수 있으므로, 후속 연구에서는 지표 정의의 다양한 후보들 사이의 대응관계를 정리하고, 설정 변화에 대한 강건성을 체계적으로 검증할 필요가 있다. 셋째, 본 논문은 정의 및 진단 기준의 제시에 초점을 두었으므로, 제안한 정의가 실제 학습 설정의 어떤 조건에서 특히 민감하게 작동하는지, 어떤 완화 기법이 어떤 구성요소($C_{dim}(t), C_{sim}(t), C_{inact}(t)$)를 우선적으로 개선하는지에 대한 실증적 분석은 후속 과제로 남는다.

ACKNOWLEDGMENT

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(15%) RS-2021-II211343(15%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

참 고 문 헌

[1] A. Kumar, R. Agarwal, D. Ghosh, and S. Levine, "Implicit under-parameterization inhibits data-efficient deep reinforcement learning," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[2] T. Moskovitz, A. K. Singh, D. J. Strouse, T. Sandholm, R. Salakhutdinov, A. D. Dragan, and S. McAleer, "Confronting reward model overoptimization with constrained RLHF," in Proc. Int. Conf. Learn. Representations (ICLR), 2024.

[3] R. D. Martinez, Z. Goriely, A. Caines, P. Buttery, and L. Beinborn, "Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing," in Proc. Conf. Empirical Methods Natural Language Process. (EMNLP), Miami, FL, USA, pp. 5999–6011, Nov. 2024.