LLM의 한국어 능력 향상에 대한 연구: Optimizer 비교 중심으로

김수민, Rahman,S M Wahidur, 조건우*, 이흥노
광주과학기술원, *조선대학교

smkim6927@gm.gist.ac.kr, sm.wahidur@gm.gist.ac.kr, *dlsrhddyd999@gmail.com, heungno@gist.ac.kr

An Improving Korean Proficiency of LLMs: A Comparative Study of Optimizers

Kim Sumin, Rahman,S M Wahidur, Jo Geon Woo*, Heung-No Lee
Gwangju Institute of Science, Technology, *Chosun University

요 약

대규모 언어모델을 새로운 언어 또는 도메인으로 확장하는 과정에서 기존 모델의 성능 저하를 억제하고 학습되지 않은 언어나 도메인의 성능을 향상시키는 것이 중요한 과제가 되었다. 본 논문은 옵티마이저 관점에서 언어모델의 새로운 언어와 도메인에 대한 성능 향상과 기존 언어 능력 성능 저하 문제를 다룬다. 이를 위해 동일 모델을 대상으로 언어-도메인이 급격하게 전이되는 학습과정에서 다양한 옵티마이저를 비교 분석하고, 각 단계의 성능 변화와 전이 지표를 통해 옵티마이저 선택이 성능 유지와 전이에 미치는 영향을 정량적으로 평가한다. 실험 결과는, 옵티마이저의 설계 선택이 단순한 학습 효율을 넘어 도메인, 언어 확장 과정에서 성능 보존과 전이에 중요한 역할을 함을 보여준다. 본 연구는 연속 사전학습 과정에서 다국어 및 도메인 확장 맥락에서 업데이트 규칙 자체를 핵심 제어 요소로 고려할 필요성을 제시한다.

Ⅰ. 서 론

대규모 언어모델을 다른 언어나 도메인으로 확장할 때는, 목표 언어 성능을 끌어올리는 과정에서 기존 영어 능력이 약화되는 문제가 반복적으로 보고된다. 이러한 현상은 사전학습을 통해 새 도메인· 언어 데이터에 적응할 때 두드러진다. 모델이 학습하지 않은 도메인 적응을 위한 추가 사전학습이 모델의 새 도메인 문제해결 성능을 개선한다는 점은 널리 확인되어 왔다. 그러나 이 과정에서 이전 지식의 손실을 동반할 수 있다[1]. 다국어 사전학습에서도 교차언어 전이를 통해, 언어 수가 늘어날수록 표현 성능이 저하되는 상충관계가 발생할 수 있다[2]. 따라서 새로운 도메인과 한국어 능력 향상을 목표로 한 추가 학습에서는 새 도메인 과 언어에 대한 적응과 기존 언어 능력 유지를 함께 고려해야 한다.

본 논문에서는 모델 수정이나 별도 모듈을 추가하지 않고, 옵티마이저로 한국어 성능을 개선하면서도 영어 능력 저하를 완화하는 가능성을 검토한다. 언어-도메인이 바뀌는 학습 과정에서 지식의 전이 양상과 성능 보존에 어떠한 차이를 만드는지 실험적으로 분석한다.

Ⅱ. 본론

사전학습된 대규모 언어모델에 한국어 데이터를 추가 학습시키는 접근법은 일반적으로 한국어 성능의 향상 방안으로 논의되어 왔다. 그 과정에서 데이터셋은 단순하게 언어만 바뀌는 것이 아닌 언어와 도메인이 함께 변화하는 복합적인 양상을 띤다.

본 연구에서는 연속 학습 환경에서 옵티마이저 선택이 전이 및 망각 특성에 미치는 영향을 분석하기 위해, AdamW, Sophia, 그리고 UPGD 를 비교했다. AdamW 는 대규모 언어모델 학습에서 표준적으로 사용된다[4]. Sophia 는 곡률 정보를 근사적으로 활용하는 방법을 이용해 제안되었다[5]. UPGD 는 연속 학습에서 음의 전이를 현상을 의미하는 plasticity 감소와 catastrophic forgetting 문제를 완화하기 위해 제안된 방법이다[6].

연속 사전학습 환경에서 전이 품질을 평가하기 위해서는 단일 단계의 성능과 과제 간 지식의 전달 양상을 함께 고려해야 한다. 성능 평가를 위해 연속학습 평가 방법인 Backward Transfer (BWT), Forward Transfer (FWT)을 이용한다. BWT 와 FWT 는 T 개의 과제가 주어질 때, 학습 과정의 성능 행렬을 $R \in \mathbb{R}^{T\times T}$, $R_{i,j}$는 i번째 과제까지 학습한 직후 j번째 과제의 테스트 성능을 의미한다.

BWT 는 학습이 진행됨에 따라 과거 과제에 대한 성능이 어떻게 변화했는지를 나타낸다.

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(R_{T,i} - R_{i,i}\right)$$

여기서 $R_{T,i} - R_{i,i} < 0$일 경우 과거 과제에 대한 성능의 하락과 지식에 대한 망각이 발생했음을 의미한다.

FWT 는 이전 단계의 학습이 미래 과제의 초기 성능에 미치는 영향을 측정한다.

$$\mathrm{FWT} = \frac{1}{T-1}\sum_{i=2}^{T}\left(R_{i-1,i} - b_i\right)$$

여기서 $R_{i-1,i}$ 는 i 번째 과제를 학습하기 직전의 성능이며, b_i는 해당 과제의 초기 기준 성능이다.

다음은 실험환경 설정이다. 본 연구에서는 생성 및 요약 과제의 성능 측정($R_{i,j}$)을 위해 ROUGE-L 과 BLEU 점수를 사용한다. 실험에 사용된 모델은 Llama3.2 로 3B

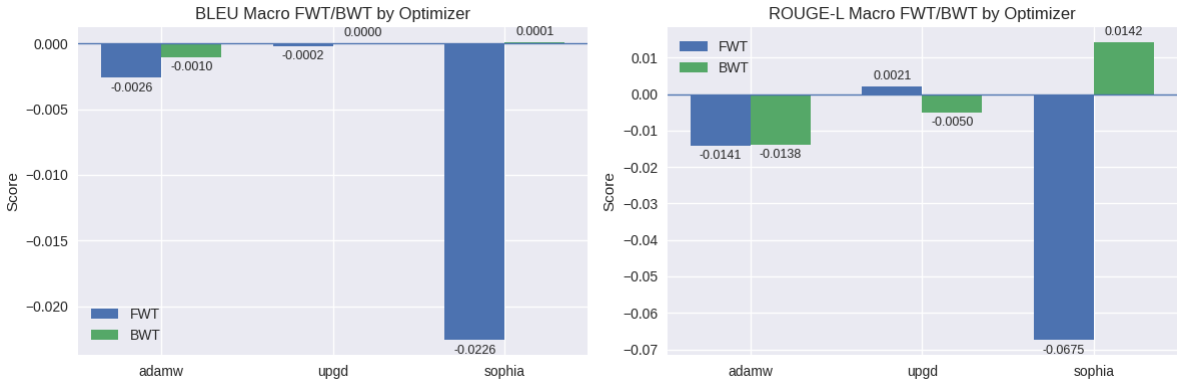

그림 1 급격한 도메인 및 언어 변화 환경에서 옵티마이저에 따른 Blue 와 Rouge-L 에 따른 BWT, FWT 의 평균값

크기이다[7]. 옵티마이저 비교 분석을 위해 3 가지의 다른 도메인 또는 언어의 데이터 셋을 다룬다. Legal Case Summarization (법률 요약)과 GSM8K (수학 추론) 과제 그리고 Korean Medical MCQA (한국어 의료 질의응답) 데이터셋이다.

비교 실험 결과에 대한 분석은 다음과 같다. AdamW 옵티마이저가 FWT 에서 Blue 는 -0.0026 을 Rouge-L 은 -0.0141 이라는 음의 값을 기록했다. 이는 전반적으로 소폭이나 음의 값을 기록했음으로 도메인이나 언어가 바뀌는 환경에서 모델이 과제 성능에 지식 전이가 활발하게 일어나지 않는 것으로 해석할 수 있다. BWT 기록을 살펴보면 Blue 는 -0.0010 을 Rouge-L 은 -0.0138 이라는 수치를 통해 과거 문제해결 능력이 새로운 지식을 학습하면서 망각현상이 일어났음을 확인할 수 있다. 불필요한 파라미터에 노이즈를 주입해 학습을 촉진시키는 UPGD 옵티마이저의 수치를 살펴보겠다. FWT, BWT 모두 Blue 에서는 -0.0002 와 0 으로 과거의 과제에 대한 성능을 크게 잊어버리지 않고 지식전이도 활발이 일어나지 않아 보이나 Rouge-L 로 보는 성능에서는 FWT 가 0.0021 BWT 가 -0.0050 으로 미세하지만 지식 전이와 망각현상이 동시에 일어난 것으로 나타난다. 2 차 근사를 이용하는 옵티마이저인 Sophia 의 수치를 분석해보면 FWT 수치가 Blue 는 -0.0226, Rouge-L 는 -0.0675 로 지식 전이가 앞의 두 옵티마이저에 비해 떨어지는 것이 확연하게 나타났다. 그러나 BWT 의 Blue 는 0.0001 로 미세하나 Rouge-L 은 0.0142 라는 양수 값을 보였다. 이는 전반적으로 새로운 언어나 도메인을 학습과정에서 성능 저하가 이뤄지지 않고 과거 도메인 또는 언어 지식이 보존되고 있음을 시사한다.

전반적인 BWT 와 FWT 를 통한 수치 분석으로 볼 때, AdamW 는 지식 흡수와 망각에서 모두 불리했다. 그러나, UPGD 가 새로운 지식을 흡수하는 데 상대적으로 큰 망각없이 안정적이며 Sophia 는 과거 지식을 보전하는 데 이점이 있으나 FWT 수치를 보면 다른 옵티마이저에 비해 새로운 지식전이에는 불리하다는 것으로 해석된다.

Ⅲ. 결론

본 논문은 영어 중심 사전학습 LLM 을 한국어로 확장하는 과정에서, 기존 영어 능력 손실을 최소화하면서 한국어 성능을 개선하는 문제를 다룬다. 핵심은 옵티마이저 차원의 접근법으로 모델의 학습에 미치는 영향을 분석하는 데 있다. 실험 결과를 분석했을 때, 한국어 적응에서도 안정적 전이를 목표로 한 업데이트 규칙에 대한 설계의 필요성을 뒷받침한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2026-RS-2021-II211835) 그리고 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (RS-2025-22932973)

We appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

참 고 문 헌

[1] Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[2] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.

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

[4] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

[5] Liu, Hong, et al. "Sophia: A scalable stochastic second-order optimizer for language model pre-training." arXiv preprint arXiv:2305.14342 (2023).

[6] Elsayed, Mohamed, and A. Rupam Mahmood. "Addressing loss of plasticity and catastrophic forgetting in continual learning." arXiv preprint arXiv:2404.00781 (2024).

[7] Grattafiori, Aaron, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).