대규모 언어 모델에서 차별적 계층 식별 방법

곽지호, 정유철*
국립금오공과대학교
wlgh6709@kumoh.ac.kr, *jyc@kumoh.ac.kr

Discriminative layer Identification Methods in Large Language Models

Jiho Gwak, Yuchul Jung*
Kumoh National Institute of Technology

요 약

본 연구는 언어 모델의 레이어별 임베딩이 서로 다른 표현적 특성을 갖는다는 점에 착안하여, 다운스트림 과제에 최적화된 레이어를 선택하는 방법을 제안한다. 선행 연구에서는 중간 레이어가 최종 출력 레이어보다 의미적·구문적 정보를 풍부하게 담는 경향이 있음이 보고되었으나, 기존 접근은 레이어 선택 과정이 체계적으로 정의되지 않았다는 한계를 가진다. 이를 보완하기 위해 본 연구는 각 레이어의 판별력을 정량적으로 측정할 수 있는 평가 과정을 설계하고, 두 개의 텍스트 분류 데이터셋을 통해 그 효과를 검증하였다. 실험 결과, 제안된 방법은 항상 최종 레이어보다 더 높은 성능을 보이는 레이어를 사전에 선택할 수 있었으며, 이는 레이어 인지적 임베딩 선택이 모델의 표현 학습을 실질적으로 향상시킬 수 있음을 보였다.

Ⅰ. 서 론

대규모 언어 모델(LLM)의 임베딩은 다양한 자연어 처리 과제에서 널리 활용되며, 보편적으로 최종 레이어의 출력이 사용된다. 그러나 최근 연구들은 모델의 각 레이어가 상이한 표현적 특성을 가지며, 특정 과제에서는 최종 레이어보다 중간 레이어가 더 풍부한 의미적·구문적 정보를 담을 수 있음을 보여주었다[1]. 이는 레이어 선택이 다운스트림 성능에 직접적인 영향을 미칠 수 있음을 시사한다.

특히 모델 파라미터를 직접 업데이트하는 전이학습이 제한적이거나 불가능한 경우에도, 사전학습된 언어 모델의 임베딩만으로도 텍스트 분류와 같은 다운스트림 과제에서 충분히 높은 성능을 확보할 수 있다는 점에서 레이어별 특성에 대한 이해는 더욱 중요하다[2]. 그러나 임베딩의 표현력을 직접 분류 성능으로 평가하는 방식은 시간적 비용이 크고, 데이터셋마다 최적의 레이어가 달라질 수 있다는 한계를 가진다.

이에 본 연구는 레이어별 임베딩의 판별력을 정량적으로 분석하는 방법을 통해, 과제 최적화에 유리한 레이어를 사전에 식별할 수 있는 가능성을 탐구한다.

Ⅱ. 본론

가. 제안 기법

본 연구에서는 언어 모델 각 레이어에서 추출한 임베딩이 가지는 판별력을 정량적으로 평가하기 위하여 실루엣 계수(Silhouette coefficient)[3]를 사용하였다. 실루엣 계수는 동일 클래스 내에서의 응집도와 클래스 간의 분리도를 동시에 고려할 수 있는 지표로, 레이어별 표현의 질을 비교하는 데 적합하다.

특정 샘플 i에 대하여 $a(i)$는 샘플 i와 동일한 클래스에 속한 다른 샘플들과의 평균 거리(클래스 내부 거리)를 의미하며, $b(i)$는 샘플 i와 가장 가까운 다른 클래스에 속한 샘플들과의 평균 거리(최근접 클래스 간 거리)를 의미한다. 실루엣 계수 $s(i)$는 $-1 \le s(i) \le 1$ 범위를 가지며, 1에 가까울수록 샘플이 같은 클래스 내에서 잘 응집되어 있고 다른 클래스와 명확히 구분됨을 의미한다. 반대로 0값에 가까우면 경계에 위치한 샘플임을, 음수일 경우 잘못된 군집에 속했을 가능성을 나타낸다. 이때 각 개별에 대한 실루엣 계수는 다음과 같이 계산된다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad ...(1)$$

데이터셋 전체에 대한 평균 실루엣 계수(S)는 다음과 같이 정의된다.

$$S = \frac{1}{N}\sum_{i=1}^{N} s(i) \quad ...(2)$$

여기서 N은 전체 샘플 수를 의미한다. 이론적인 측면에서는 평균 실루엣 계수가 높을수록 분류 정확도가 향상될 가능성이 크지만, 임베딩의 형태만으로 분류 성능을 직접적으로 보장하기는 어렵다. 따라서 본 연구에서는 실루엣 계수의 절대값보다는 레이어별 계수의 형태와 상대적 변화율의 변곡 지점에 주목하여, 판별력이 높은 레이어를 식별하였다..

나. 실 험

본 연구의 실험 설정은 다음과 같다. 제안한 방법의 효과를 검증하기 위하여 두 개의 성격이 다른 텍스트 분류 데이터셋을 사용하였다. 첫 번째는 AG News[4]로, 4개의 클래스로 구성된 뉴스 기사 주제 분류 데이터셋이며, 정치, 경제, 과학·기술, 스포츠와 같은 다양한 도메인을 포함한다. 두 번째는 Emotion[5]으로, 6개의 클래스를 가진 문장 단위 감정 분류 데이터셋이며, joy, sadness, anger, fear, surprise, love 등의 감정 레이블을 포함한다. 두 데이터셋은 서로 다른 언어적 특성과 난이도를 가지므로, 제안된 방법이 다양한 조건에서 일관적으로 성능을 보일 수 있는지를 평가하는 데 적합하다.

모델은 두 가지 Decoder 기반 대규모 언어 모델을 실험에 활용하였다. 하나는 Qwen3-1.7B[6]로 총 26개의 레이어를 가진 중형 규모 모델이며, 비교적 깊은 구조를 통해 풍부한 계층적 표현을 제공한다. 다른 하나는 LLaMA-3.2-1B[7]로 총 16개의 레이어를 가지며, 경량 구조임에도 불구하고 최신 아키텍처의 강점을 활용할 수 있다. 서로 다른 크기와 레이어 수를 가진 두 모델을 병행하여 실험함으로써, 제안된 방법이 모델 규모나 구조에 관계없이 적용 가능함을 확인하기 위함이다.

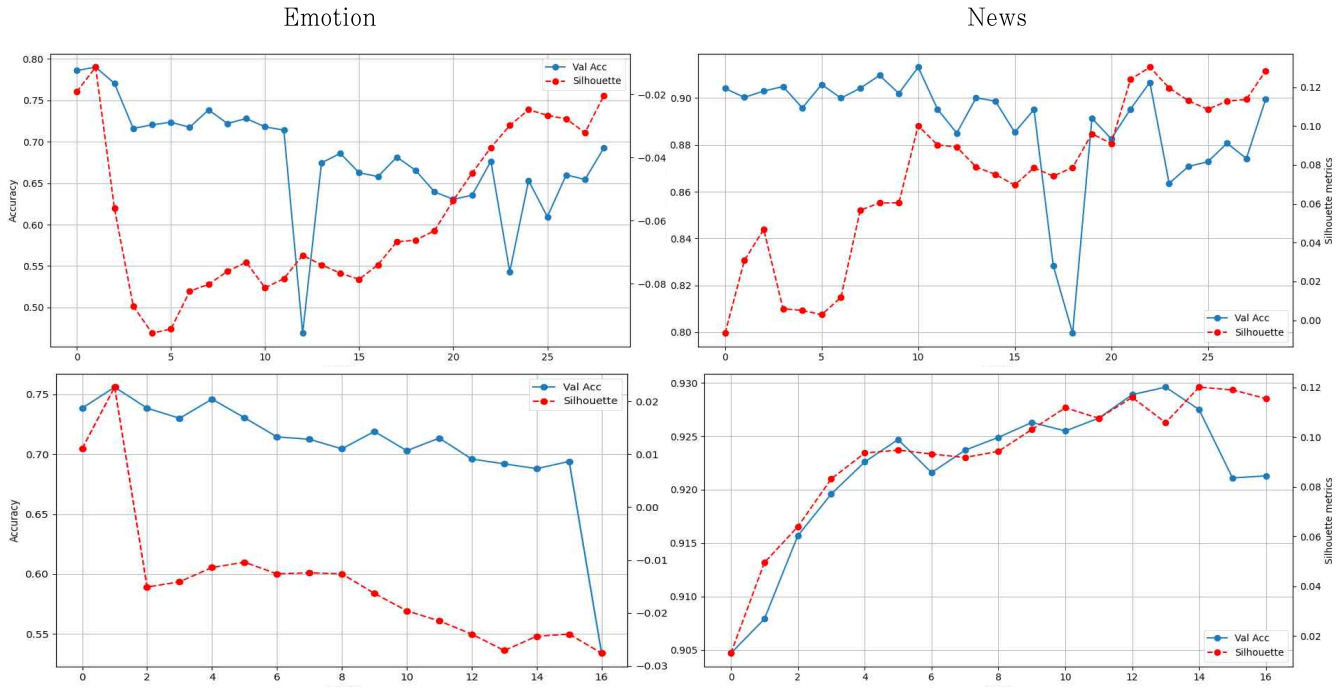

그림 1 각 모델에 대한 레이어별 정확도와 평균 실루엣 계수

실험 방법은 다음과 같다. 각 모델로부터 레이어별 임베딩을 추출한 뒤 평균 실루엣 계수를 산출하여 판별력을 비교하였다. 더 나아가 단순 Linear Layer를 활용한 분류 성능을 함께 평가하고, 실루엣 계수의 변곡점과 상대적 변화 추세를 관찰함으로써 사전에 최적의 레이어를 식별할 수 있는 가능성을 탐구하였다.

다. 결 과

그림 1은 각 모델과 데이터셋에 대해 레이어별 분류 정확도와 평균 실루엣 계수의 변화를 함께 나타낸 것이다. 상단의 두 그래프는 Qwen-3 1.7B 모델의 결과이며, 하단의 두 그래프는 LLaMA-3.2-1B 모델의 결과를 보여준다. 각 그래프는 좌측이 Emotion 데이터셋, 우측이 AG News 데이터셋을 대상으로 한 실험이다.

실험 결과에서 확인할 수 있는 주요 시사점은 두 가지이다. 첫째, 전반적으로 분류 정확도의 추세가 평균 실루엣 계수와 유사하게 나타났다. 이는 실루엣 계수가 단순한 군집 품질 지표를 넘어 실제 분류 성능과 일정 수준의 상관성을 가진다는 점을 보여준다. 둘째, 평균 실루엣 계수의 변곡점에 해당하는 구간에서 대부분의 경우 분류 정확도가 가장 높은 값을 기록하였다. 이는 단순히 최종 레이어에 의존하기보다는, 중간 레이어가 판별력 있는 표현을 담고 있음을 시사한다.

흥미로운 점은 두 모델과 두 데이터셋 모두에서 이러한 경향이 일관되게 나타났다는 것이다. 특히 감정 분류 태스크의 경우 두 모델 모두 극 초반 레이어에서 가장 높은 성능을 보였는데, 이는 짧고 직관적인 문장 단위 감정 분류 과제가 비교적 얕은 계층의 표현만으로도 효과적으로 수행될 수 있음을 보여준다. 반면 뉴스 주제 분류에서는 중간 레이어에서 뚜렷한 변곡점이 관찰되었으며, 이는 보다 복잡한 의미적 구분에 중간 수준 표현이 중요함을 나타낸다. 모델 규모(26-layer Qwen vs. 16-layer LLaMA)와 데이터셋 성격(뉴스 주제 분류 vs. 감정 분류)이 상이함에도 불구하고, 실루엣 계수의 상대적 변화 추세는 중요한 레이어를 식별하는 근거로 작동하였다. 이는 본 연구에서 제시한 분석 방법이 특정 모델이나 특정 데이터셋에 종속되지 않고, 다양한 환경에서 일반적으로 활용될 수 있음을 시사한다. 표 1은 모델, 데이터셋에 따른 평균 실루엣 계수와 정확도의 변화 추이와 변곡점을 나타낸 표이다.

	Dataset	변화 추이		변곡점(정확도)
		실루엣	정확도	
Qwen	Emotion	↘	↘	**1(0.7900)**, 24
	AGNews	→	↘	2, **10(0. 9132)**, 22
Llama	Emotion	↘	↘	**1(0.7560)**, 5
	AGNews	↗	↗	12(0.9289), 14

표 1 평균 실루엣 계수, 정확도 변화 추이와 변곡점

	Dataset	최대 정확도	및 레이어
Qwen	Emotion	0.7900	1
	AGNews	0.9132	10
Llama	Emotion	0.7560	1
	AGNews	0.9296	13

표 2 최대 정확도 및 레이어

Ⅲ. 결론

본 연구에서는 대규모 언어 모델의 레이어별 임베딩이 가지는 판별력을 정량적으로 분석하고, 이를 통해 중요한 레이어를 식별할 수 있는 가능성을 탐구하였다. 실루엣 계수를 활용하여 레이어별 응집도와 분리도를 평가하였으며, 단순 분류 성능과 비교한 결과, 실루엣 계수의 상대적 변화와 변곡점이 분류 정확도와 일정 수준의 상관성을 가지는 것을 확인하였다. 결론적으로, 본 연구는 레이어별 실루엣 계수 분석이 중요한 레이어를 사전에 판별하는 근거로 활용될 수 있음을 보여주었으며, 이는 fine-tuning이 제한적인 환경에서 임베딩 기반 학습을 위한 실질적 대안이 될 것으로 기대된다. 향후 연구에서는 보다 다양한 언어 모델과 태스크에 본 접근을 확장 적용하고, 단순 레이어 선택에 그치지 않고 Attention 메커니즘을 포함한 복합적인 평가 기법을 활용함으로써, 중요한 표현을 더욱 정교하게 식별할 수 있는 방법론을 모색할 예정이다. 이를 통해 본 연구의 접근법을 일반화된 레이어 판별 프레임워크로 발전시키는 것이 목표이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 - 학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임 (IITP-2025-RS-2022-00156394)

참 고 문 헌

[1] Ju T., Sun W., Du W., Yuan X., Ren Z., Liu G. "How Large Language Models Encode Context Knowledge? A Layer-Wise Probing Study," arXiv preprint arXiv:2402.16061

[2] Skean O., Arefin M. R., Zhao D., Patel N., Naghiyev J., LeCun Y., Shwartz-Ziv R. "Revisiting Layer-Wise Information in Language Models," arXiv preprint arXiv:2502.02013.

[3] Rousseeuw, P.J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53 - 65, 1987.

[4] Zhang X., Zhao J., LeCun Y. "Character-level Convolutional Networks for Text Classification," Advances in neural information processing systems 28, pp. 649-657.

[5] Saravia E., Liu H. C. T., Huang Y. H., Wu J., Chen Y. S. "Carer: Contextualized affect representations for emotion recognition," Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 3687-3697.

[6] Bai J., et al. "Qwen Technical Report," arXiv preprint arXiv:2309.16609.

[7] AI @ Meta. "The Llama 3 Herd of Models," arXiv preprint arXiv:2404.11225.