102 희소 어텐션

Sparse Attention

모든 입력 대신 핵심 정보에만 집중해 효율을 높이는 AI 연산 구조

- 문장 전체의 단어를 비교하지 않고, 의미상 중요한 일부 관계에만 주의를 집중해 연산량을 줄이는 어텐션 기법
- 대규모 입력 데이터를 빠르고 효율적으로 처리하기 위해 설계된 경량화된 AI 연산 방식

희소 어텐션의 개념

희소 어텐션은 AI의 핵심 구성 요소인 어텐션 구조에서 발전한 기술로, 방대한 입력 중 중요한 정보에만 선택적으로 집중해 연산 효율을 높이는 방식입니다. 어텐션은 문장 내 단어 간 관계를 분석해 의미를 파악하는 메커니즘으로, 사람이 문장을 읽을 때 핵심 단어에 주의를 기울이는 과정과 유사합니다. 기존의 완전 어텐션(full attention)은 모든 단어와 토큰간 관계를 계산해야 하므로 문장이 길어질수록 계산량과 메모리 사용이 폭증해 대형 모델에서는 비효율이 컸습니다. 희소 어텐션은 이러한 문제를 해결하기 위해 핵심 연결만 남기고 불필요한 계산을 생략해, 필요한 부분에만 주의를 집중하도록 설계되었습니다.

희소 어텐션의 작동 방식

희소 어텐션의 핵심 원리는 "선택적 집중"입니다. 입력된 데이터 중 중요도가 높은 요소만 선별해 관계를 계산하고, 나머지는 생략합니다. 이 선택은 주로 세 가지 방식으로 이루어집니다. 첫째, 문장 내 인접한 단어들끼리만 관계를 계산하는 인접 관계 중심 방식, 둘째, 문장 전체를 살피되 핵심 단어에 주의를 집중하는 의미 중심 방식, 셋째, 두 방식을 결합해 효율과 정확성의 균형을 맞추는 혼합형 구조입니다. 문장 길이가 길어져도 계산량이 일정하게 유지되어 긴 문서나 시계열 데이터 처리에 매우 효과적입니다.

희소 어텐션의 중요성

희소 어텐션은 필요한 부분만 선택적으로 계산해 연산을 줄이는 구조로, 전문가 조합(MoE)의 '일부 전문가만 활성화'하는 희소성 원리와 맥락을 같이합니다. MoE가 필요한 전문가만 골라 연산 부담을 줄이듯, 희소 어텐션도 입력 토큰 중 핵심 정보만 계산해 효율성을 극대화합니다. 이를 통해 긴 문장·대용량 텍스트·영상·음성 등 복잡한 데이터를 빠르고 정확하게 처리할 수 있으며, 과거보다 긴 토큰을 사용할 수 있습니다. 대표적인 예로 Longformer(AI2), BigBird(Google), Sparse Transformer(OpenAI) 등이 있으며, 다양한 LLM과 생성형 AI에도 적용되고 있습니다. 희소 어텐션은 속도, 비용, 에너지 효율을 모두 개선해 AI가 복잡한 정보를 사람처럼 이해할 수 있도록 돕는 지능형 연산 최적화 기술이며, 향후 모델 경량화와 친환경 AI 구현의 핵심 기반이 될 것입니다.