LLaMA Adapter에 기반한 문서 그라운딩된 응답 생성

LLaMA Adapter for Document-Grounded Response Generation

이 검○[1], 최형준[2], 홍범석[3], 최원석[4], 한영섭[5], 전병기[6], 나승훈[7]

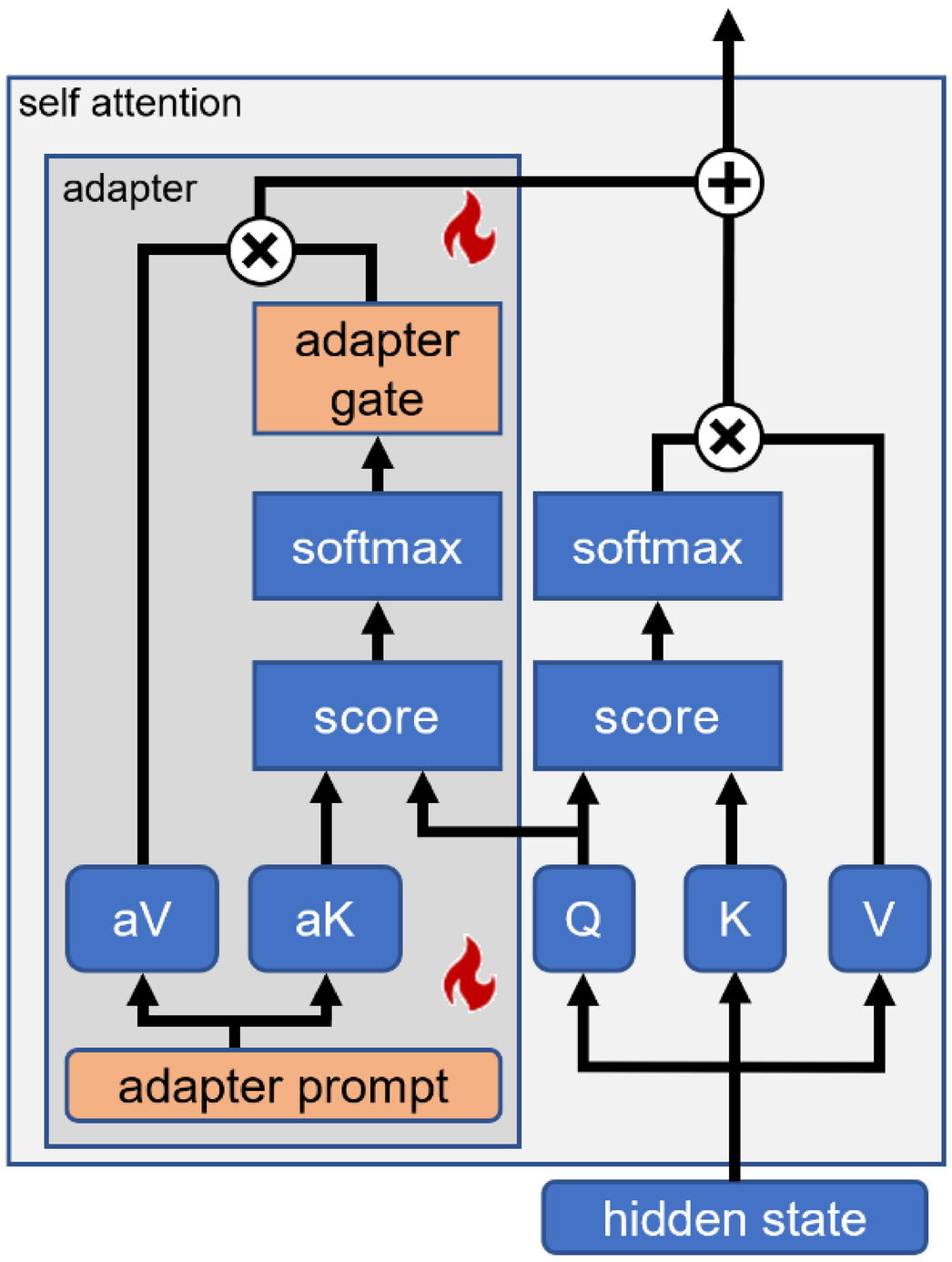
전북대학교[127], LG 유플러스[3456]

Ⅰ. 서론

연구 목적

문서 기반 대화 시스템에서 답변 생성의 성능 개선은 매우 중요한 부분이다. 하지만 생성에 사용되는 사전 학습된 언어 모델을 Fine-tuning하기 위해서는 많은 연산과 시간이 필요하다. 본 연구에서는 LLaMA-Adapter 논문에서 제안된 Adapter 구조를 실험한다. 사전 학습된 LLaMA 7B 모델의 파라미터는 Freeze하고, 추가한 LLaMA-Adapter만 학습하는 방식이 답변 생성 과정에 효과적인지 문서 기반 대화 시스템의 학습에 사용되는 MultiDoc2Dial 데이터 셋을 사용하여 실험하고 성능을 확인한다.

Ⅱ. 실험 설계



LLaMA-Adapter 방식 및 모델 구성

Adapter prompt는 [virtual token length, model hidden] 으로 초기화하고, Adapter gate는 num_head 크기로 0으로 초기화한다.

query와 adapter key의 softmax 연산이 끝난 후 adapter gate 값을 곱해준다. 0으로 초기화된 gate가 랜덤 초기화된 것에 비해 학습 초기에 adapter에 의한 noise를 제거하여, 더 높은 성능을 얻게 해준다.

adapter cross attention의 결과값은 기존 Self attention과 더하여 out projection matrix의 입력으로 사용한다.

LLaMA 7B 모델을 freeze하고 LLaMA-Adapter를 추가하여 adapter prompt와 gate만 학습한다. bf16 mixed precision을 사용하였다.

데이터 셋

문서 기반 대화 시스템의 학습을 위해 멀티 턴 대화와 문서가 포함된 MultiDoc2Dial 데이터 셋을 사용했다. 멀티 턴 대화 내에서 agent의 응답에 대해 한 개의 문서가 참조 대상으로 주어지는 것을 토대로 한 개의 멀티 턴 데이터를 [문서 - 대화 기록 - 사용자의 질문 - 응답] 형태의 여러 샘플로 분할하여 다음과 같이 특수한 텍스트를 포함하여 구성하였다.

문장 〈h〉 정답 span 〈/h〉 문장 〈l〉 대화 기록 〈/l〉 〈q〉 질문 〈/q〉 정답

전처리 과정에서 GPU 연산과 시간 효율을 위해 문서가 너무 길거나 생성할 정답이 너무 긴 학습 데이터는 제거하였고, 학습, 평가 데이터를 각각 20588개, 4056개를 사용했다.

Ⅲ. 실험 결과

표 2: 답변 생성 성능

실험 방식	layer	length	learning rate	F1	Meteor	Rouge-L	ScareBLEU
LLaMA - Adapter	10	10	1e-2	35.56	29.93	33.83	12.96
LLaMA - Adapter	10	30	1e-2	37.43	32.43	35.70	15.64
LLaMA - Adapter	30	10	1e-2	**60.62**	**58.37**	**59.32**	**47.63**
LLaMA - Adapter	30	30	1e-2	59.70	57.08	58.32	46.12
LoRA (r=1, α=1)	-	-	3e-4	59.97	57.34	58.61	46.35

결과 분석

LLaMA - Adapter를 적용한 실험들 중에서 layer 30, length 10일 때 가장 우수한 성능을 얻었고, LoRA를 Self attention Q, K, V, O matrix에 적용한 것보다 약간 더 높은 성능을 확인할 수 있었다. LoRA는 layer 30, length 30일 때의 성능보다 약간 더 높은 성능을 기록했다.

softmax의 지수 연산을 통해 지수 스케일로 변화된 softmax의 결과에 동일한 값을 곱해도, softmax에 의해 나온 결과의 top 몇 개인 값만 최종 결과에 영향이 클 것으로 생각되고, 점수 차이는 우연한 것으로 생각된다.

문서 기반 대화 시스템의 응답 생성 과정에 사용하는 모델의 layer마다 별도의 prompt를 주어 Cross attention하는 방식이 성능 향상에 큰 기여를 했음을 알 수 있다.

Ⅳ. 결론 및 향후 연구

결론

문서 기반 대화 시스템에서 LLaMA-Adapter가 LoRA를 적용한 것보다는 조금 더 높은 성능을 얻을 수 있음을 확인하였다.

결과 분석을 통해 LLaMA-Adapter 의 adapter prompt token length가 10 미만일 때에도 유사한 성능을 낼 수 있을지 추가 실험을 진행할 계획이다.

향후 연구로는 LoRA 이외의 다른 PEFT 방법론들과 비교 실험을 진행하고, 어떻게 성능 차이가 나는지 확인해볼 예정이다. 또한, LLaMA-Adapter를 사용하여 GPT계열 모델 또는 T5의 디코더 층에 적용하여 언어 모델간 성능을 비교한 뒤 한국어 모델에 적용해 볼 예정이다.