Deep Prompt Tuning 기반 한국어 질의응답 기계 독해

김주형◦, 강상우
가천대학교 AI · 소프트웨어학부
stephano12@gachon.ac.kr◦, swkang@gachon.ac.kr

Deep Prompt Tuning based Machine Comprehension on Korean Question Answering

Juhyeong Kim◦, Sang-Woo Kang
School of Computing, Gachon University

요 약

질의응답 (Question Answering)은 주어진 질문을 이해하여 그에 맞는 답변을 생성하는 자연어 처리 분야의 핵심적인 기계 독해 작업이다. 현재 대다수의 자연어 이해 작업은 사전학습 언어 모델에 미세 조정 (finetuning)하는 방식으로 학습되고, 질의응답 역시 이러한 방법으로 진행된다. 하지만 미세 조정을 통한 전이학습은 사전학습 모델의 크기가 커질수록 전이학습이 잘 이루어지지 않는다는 단점이 있다. 게다가 많은 양의 파라미터를 갱신한 후 새로운 가중치들을 저장하여야 한다는 용량의 부담이 존재한다. 본 연구는 최근 대두되는 deep prompt tuning 방법론을 한국어 추출형 질의응답에 적용하여, 미세 조정에 비해 학습시간을 단축시키고 적은 양의 파라미터를 활용하여 성능을 개선했다. 또한 한국어 추출형 질의응답에 최적의 prompt 길이를 최적화하였으며 오류 분석을 통한 정성적인 평가로 deep prompt tuning이 모델 예측에 미치는 영향을 조사하였다.

주제어: 추출형 질의응답, prompt 기반 전이학습, deep prompt tuning

1. 서론

질의응답(Question Answering)이란 사람과 질문과 대답을 주고받는 작업으로, 70년대부터 시작되어 자연어 처리 인공지능 분야에서 많은 연구가 이루어진 핵심적인 주제이다. 최근의 질의응답 연구는 다른 다양한 자연어 이해 작업과 마찬가지로 Transformer[1] 기반의 사전학습 모델을 질의응답 작업에 맞게 전이학습(Transfer Learning)[2]하는 방식을 채택하고 있다.

전이학습의 대표적인 방법으로는 사전 학습된 파라미터를 새로운 작업의 데이터셋을 사용하여 모델을 부분적으로 다시 학습시키는 미세 조정(Fine-tuning)[3]이 있다. 미세 조정은 부분 재학습을 진행하여 수행하고자 하는 작업에 특화된다는 장점이 있지만, 모델의 크기가 큰 경우 재학습에 많은 시간과 연산 비용이 소모된다는 한계가 있다. 특히 초거대 언어 모델의 경우, 미세 조정이 언어 모델의 크기에 비해 상대적으로 너무 미세하게 작용하여 전이학습의 효과가 나타나지 않는다. 자연어 분야의 최신 연구 동향이 초거대 언어 모델의 활약에 주목한다는 점에서 이러한 미세 조정의 단점들은 더욱 부각될 전망이다.

미세 조정의 단점을 해결하고자 사전학습 모델을 동결하고, 수행하려는 작업에 관한 지식을 prompt[4]로 작성하여 활용하는 prompt 기반 전이학습 방식이 논의되었다. 초기 연구는 사람이 데이터를 보고 수동으로 입력하는 prompt(manual prompt)를 사용했으나, 많은 노력을 필요로 하고 최적의 결과를 보장하지 않는다.

Prompt tuning[5]이란 모델의 파라미터는 동결하고, 작업을 수행하기 위한 별도의 갱신가능한 prompt(soft prompt)[6]를 추가하여 이를 조정하는 전이학습 방법론이다. Prompt tuning에서 학습 시작 시 prompt는 임의의 숫자 형태로 초기화되어 사전 학습 모델의 입력층의 입력 임베딩 앞부분에 결합된다. 이후 반복적으로 이뤄지는 학습의 역전파 과정에서 사전학습 모델의 파라미터는 동결되어 갱신되지 않으며, 입력된 prompt는 예측의 오차가 줄어드는 방향으로 갱신된다. 학습을 완료하여 해당 작업에 특화된 prompt는 동결된 사전학습 모델의 가중치와 결합하여 해당 작업을 위한 추론에 활용될 수 있다. Deep prompt tuning이란 prompt tuning에서 prompt 임베딩을 입력층만이 아닌 입력층부터 출력층까지 모든 층에 적용하는 방식으로 prefix tuning[7], p-tuning v2[8] 등이 이에 해당한다.

본 연구는 deep prompt tuning이 미세 조정에 비해 효율적이고 동시에 효과적이다는 [7],[8]의 결과에 주목하여, 해당 방법을 한국어 질의응답에 적용하여 실험한다. KorSQuAD 1.0 (The Korean Question Answering Dataset, 한국어 질의응답 데이터셋)[9]에 대한 실험 결과, deep prompt tuning을 적용하였을 때 연산량과 학습 시간이 단축되었으며 large 크기의 모델에서 기존 미세 조정 대비 0.3~0.5%의 성능이 향상하였다.

2. 관련 연구

2.1 추출형 질의응답

질의응답 작업 중에서 주어진 문맥의 내용을 이해하고 문맥의 내용 속에서 답변을 찾아 답변하는 작업을 추출형 질의응답 (Extractive Question Answering)[10]이라고 한다. 추출형 질의응답이란 문서 (context)와 질의(question)가 주어졌을 때, 해당 문서 내에서 텍스트 범위(spans of text)의 시작 지점과 끝 지점을 질문에 대한 답변(answer)으로 추출하는 기계 독해 작업이다. 문서와 질의, 답변의 구성 예시는 표 1.과 같으며 대표적인 데이터셋으로는 SQuAD (Standford Question Answering Dataset) 1.0[11], SQuAD 2.0[12] 등이 있다.

표 1. 추출형 질의응답 예시

문서:	"… 실증 등 연구개발 및 기반 구축을 추진하고 상용경수로에서 개발되는 안전기술을 제4세대 원전으로 확장한다. 공고해진 원자력 산업기반을 바탕으로, …"
질의:	"안전기술의 원전 확장을 위해 개발되는 것은?"
답변:	"상용경수로"

Transformer 구조의 사전학습 모델은 인코더 혹은 디코더 구조에 따라 다른 특성을 지닌다. BERT[13], RoBERTa[14] 등의 인코더 모델은 "하늘은 왜 파란색인가"와 같은 개방형 질문(open-ended questions)는 제대로 처리하지 못하지만, "누가 비행기를 발명했는가"와 같은 사실적인 질문(factoid questions)의 답변에는 우수하게 대답하는 특성을 보인다. 따라서 이러한 BERT계열 모델은 주어진 사실 내에서 답변을 추출해야 하는 본 작업에 효과적이다.

2.2 Prompt 기반 전이학습

미세 조정은 재학습 과정에서 모델의 전체 파라미터를 갱신하기 위해 많은 양의 메모리와 시간을 요구하며, 파라미터가 수십억 개 이상인 초거대 생성 모델의 경우 미세 조정이 효과적으로 이루어지지 않는다. 또한, 각 작업에 맞게 모델을 미세 조정한 경우, 갱신된 모델의 복사본을 저장해야 한다는 점에서 용량의 부담이 크다.

따라서 prompt를 활용한 전이학습이 고안되었고, 초기에는 수행하는 작업에 적합할 것으로 예상되는 자연어 형태의 고정된 prompt를 수동으로 넣는 방식으로 이루어졌다[4]. 하지만 고정된 prompt는 형태가 조금만 바뀌더라도 모델의 성능이 극단적으로 변하기 때문에 여러가지 prompt template에 대해 사람이 일일이 해봐야 한다는 단점이 있다.

Auto prompt[15]는 최초로 자동적으로 prompt를 생성하는 방식을 제안한 논문으로, 언어 모델에 있는 단어들 중 모델의 추론 결과의 오차에 따라 gradient based search를 통해 최적의 prompt를 탐색하였다. 후속 연구인 lottery prompt[16]는 단어들을 출현 빈도 순으로 정렬한 후, 영어의 어순에 맞춘 템플릿에 {명사-동사-형용사/부사/전치사} 모든 조합의 경우의 수를 만들어 최적의 prompt를 찾는 방식을 제안하였다.

자동적으로 prompt를 생성할 수 있더라도 자연어 형태로 고정하여 임베딩하는 것은 신경망의 특성과 맞지 않고 여전히 반복적인 탐색을 필요로 한다는 한계가 존재한다. 따라서 prompt 생성 방식을 신경망의 학습과 같은 원리로 연속적인 공간에서 생성하는 prompt tuning 방법들[5][7][8]이 고안되었다. [7][8]은 사전학습 모델을 동결시킨 상태에서 입력으로 들어가는 소수의 prompt만 학습시켜도 일부 자연어 이해 하위 작업(downstream task)에 대해서 미세 조정과 유사하거나 우위에 있는 성능을 낼 수 있음을 보였다.

3. Deep Prompt Tuning

Deep prompt tuning이란 입력층부터 출력층까지 모든 층에 적용하는 방식으로 prefix tuning[7], p-tuning v2[8] 등이 이에 해당한다. [7]는 prompt를 모든 은닉층에 같은 값으로 사용하지 않고 각 층마다의 prompt를 앞에 결합하는 방식으로(prefix) 추가한다. [7]은 추가된 prompt 토큰들이 하이퍼볼릭 탄젠트 함수가 결합된 다층 퍼셉트론 (Multi-Layer-Perceptron)을 통과한 후 임베딩이 되도록 구현하였다. 하지만 p-tuning v2[8]에서 추출형 질의응답과 유사한 논리 질의응답 (Boolean Question Answering)에서 다층 퍼셉트론이 prompt 파라미터를 학습하는데 소음으로 작용하기 때문에 prompt를 바로 임베딩하는 것이 성능이 좋다고 밝힌 바 있다.

따라서 본 연구의 모델 구조는 사전학습 모델을 완전히 동결시킨 후, 각 은닉층 앞에 조정가능한 prompt 토큰들을 바로 임베딩하는 [8]의 구조를 적용하여 추출형 질의응답을 수행하였으며 그림 1과 같다. 먼저 추출형 질의응답을 위한 BERT 계열 사전학습 모델의 input_ids에는 [CLS] 토큰 다음에 토큰화된 질의, 토큰화된 문서가 차례대로 이어 붙어 임베딩된 후 입력된다. L개의 은닉층을 가진 사전학습 모델에 의해 학습되는 길이 N인 prompt $P = \{p_{00}, p_{01}, \ldots, p_{LN}\}$ 는 학습 시작 시 임의의 값으로 초기화되어(각 prompt 토큰이 해당 계층에서의 순서 위치값으로 초기화하였다) 각 층의 앞부분에 결합하여 임베딩된다. 마지막 은닉층의 출력은 선형 계층과 SoftMax 함수를 포함한 최종 출력층으로 공급되어

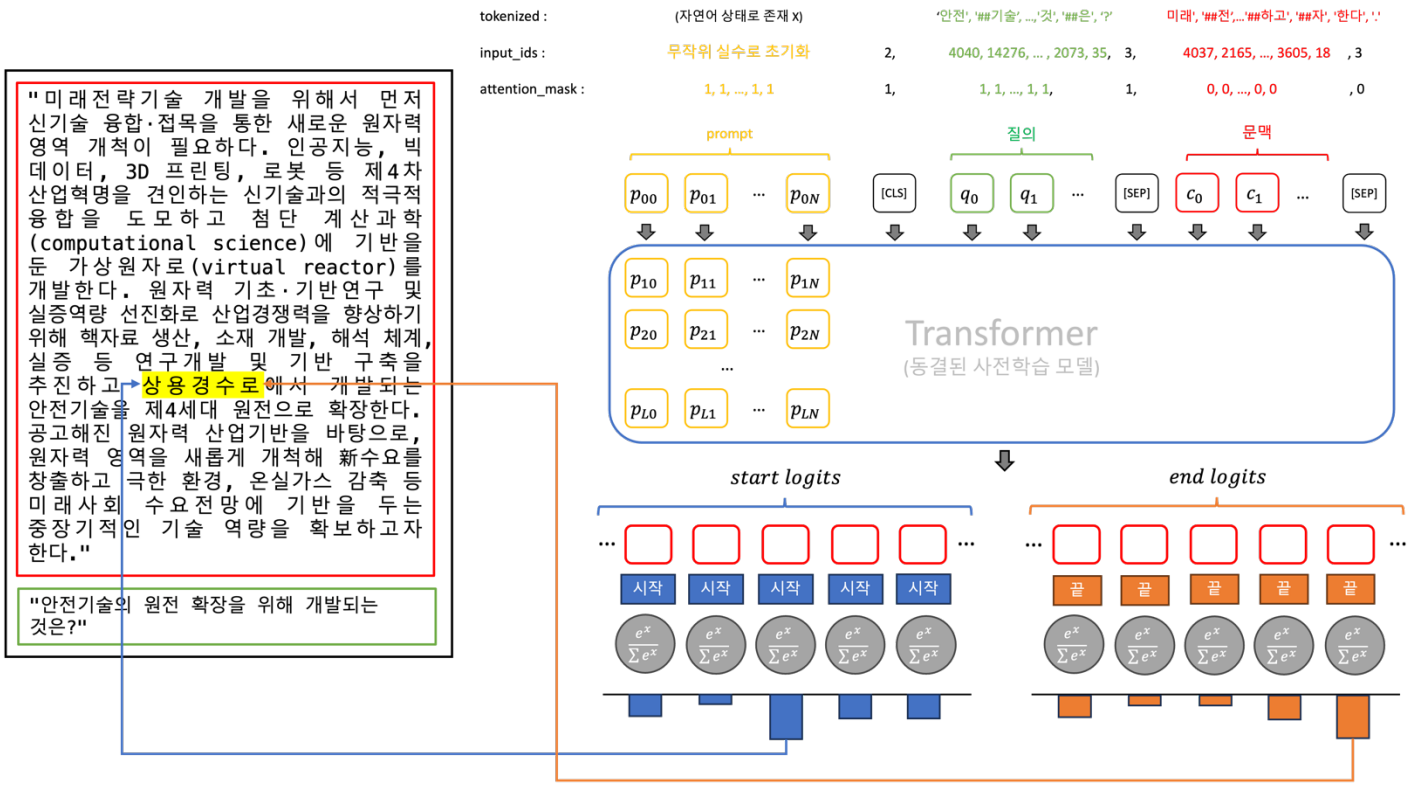


그림 1 Deep Prompt Tuning을 적용한 추출형 질의응답 구조

정답의 시작 지점과 끝 지점에 대한 확률을 logits으로 출력한다. 입력 시퀀스의 길이를 I, 은닉층의 출력을 $H = \{h_1, h_2, \dots h_I\}$라고 할 때, 시작 지점의 벡터 S와 끝 지점의 벡터 E에 대한 확률 $Pr_{start}(i)$ 와 $Pr_{end}(i)$ 은 수식 (1), (2)와 같다.

$$Pr_{start}(i) = \frac{\exp(S \cdot h_i)}{\sum_{i=0}^{I} \exp(S \cdot h_i)} \quad \cdots (1)$$

$$Pr_{end}(i) = \frac{\exp(E \cdot h_i)}{\sum_{i=0}^{I} \exp(E \cdot h_i)} \quad \cdots (2)$$

계산된 시작 위치와 끝 위치를 cross entropy 손실 함수를 이용하여 실제 정답 시작 위치와 끝 위치와의 오차를 계산한다. 계산된 시작 위치 오차와 끝 위치 오차의 평균값을 모델의 전체 프롬프트 토큰 $P = \{p_{00}, p_{01}, \dots, p_{LN}\}$에 역전파하여 갱신한다.

4. 실험 및 평가

4.1 데이터셋

KorQuAD 1.0. KorQuAD (The Korean Question Answering Dataset, 한국어 질의응답 데이터셋)[9]은 LG CNS에서 구축한 대규모 질의응답 데이터셋으로 한국어 기계독해 모델을 학습하고 그 성능을 평가하는데 사용되고 있다. KorQuAD는 1.0과 2.0 두 가지 버전이 존재하며 본 연구는 1.0 버전으로 실험을 진행하였다. KorQuAD 1.0의 전체 데이터는 1,560 개의 위키피디아 문서에 대해 10,645 문단과 66,181 개의 질의응답 쌍으로 구성되어 있다. 66,181개의 질의응답은 학습용 데이터 60,407개와 평가용 데이터 5,774개로 나뉜다. 데이터의 구성 형식은 SQuAD 1.0[11], SQuAD 2.0[12]과 완전히 같으며, 평가 방식 역시 마찬가지로 완전 일치 (Exact Match, 모델이 정답을 정확히 맞춘 비율)와 F_1 점수 (F_1 score: 모델이 난 답안과 정답을 음절 단위로 비교하여 정답을 겹치는 부분을 고려한 부분 점수)의 척도로 성능을 수치화한다.

4.2 실험 환경

본 실험은 KLUE (Korean Language Understanding Evaluation, 한국어 이해능력 평가) 벤치마크[17]를 통해 사전 학습된 KLUE/BERT-base, KLUE/RoBERTa-base, KLUE/RoBERTa-large을 사용하였다. 이 모델들은 모두 자연어 이해 작업에 특화 되어있는 양방향 인코더 모델로, base 크기의 두 모델은 12개의 은닉층과 크기 768의 은닉 상태, 110M개의 파라미터로 구성 되어있고,

모델 (KLUE)	F_1 점수		완전 일치		파라미터 개수		총 학습시간 (분)	
	FT	DPT	FT	DPT	FT	DPT[1]	FT	DPT
BERT-base	90.58	86.64	85.43	80.20	110M	331,776	98.91	67.15
RoBERTa-base	91.57	90.03	86.61	84.32	110M	259,586	95.24	73.71
RoBERTa-large	92.02	**92.36**	86.89	**87.37**	336M	690,178	275.18	206.16

표 2. 성능 비교 및 학습 소요 비교 (FT: fine-tuning, DPT: Deep Prompt Tuning)

large 크기의 모델은 24개의 은닉층과 크기 1024의 은닉 상태, 340M개의 파라미터로 구성 되어있다.

본 실험은 추출형 질의응답에서 각 질문에 대해 몇 개의 정답 후보지를 생성할 것인지를 결정하는 인자 N을 20으로 설정하였다. 즉, 모델은 질문이 주어졌을 때 본문에서 20개의 정답 부분을 예측한 후, 이 중 가장 확률이 높은 하나의 정답 후보를 선택하고 그 선택된 답변에 대해 평가가 이루어진다. 예측 후보지를 20개 생성할 경우 실제 정답이 이 중에 포함되어, 실제 정답을 맞추지 못하더라도 실제 정답에 대한 확률을 확인할 수 있다.

학습은 각 모델에 대해 각각 3 에폭 씩 시행하였으며, 배치 사이즈의 경우 base 크기의 모델은 32, large 크기의 모델은 16으로 진행하였다. 학습률은 미세 조정의 경우 2e-5, Deep prompt tuning의 경우 5e-3으로 진행하였다. GPU 자원은 Nvidia Titan RTX 2기를 사용하였다.

4.3 결과

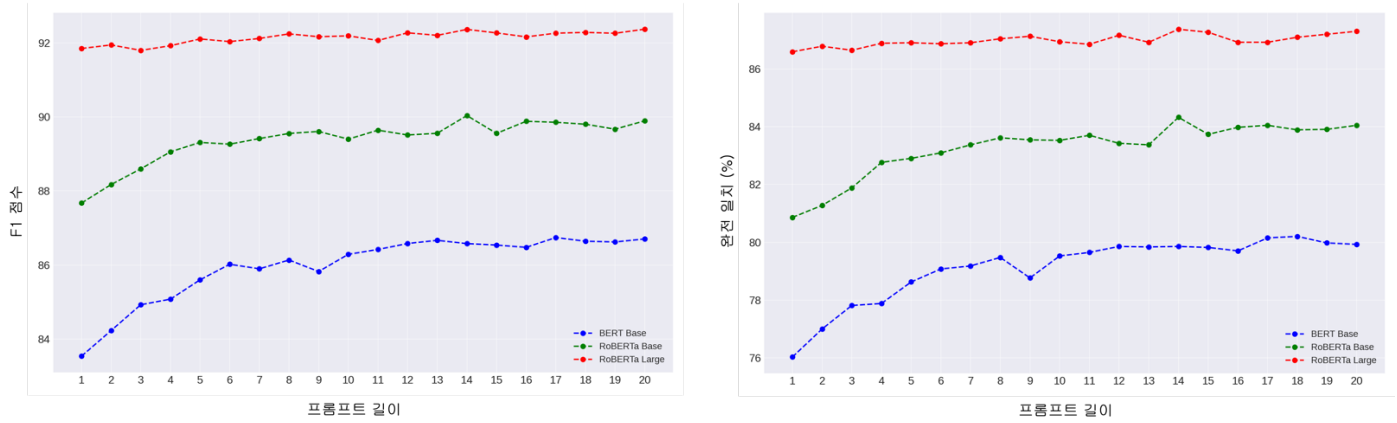
표 2는 각각 미세조정과 deep prompt tuning의 성능과 발생하는 연산적 소요인 파라미터 개수와 총 학습시간을 나타낸다. Deep prompt tuning의 학습 파라미터 개수는 사전학습 모델 대비 0.3~0.2%로 미세 조정에 비해 현저히 적은 것을 확인할 수 있었다. 총 학습시간의 경우,기존의 미세 조정에 비해 BERT-base, RoBERTa-base, RoBERTa-large 각각에 대해 32%, 22%, 25% 학습시간이 단축되었다. 성능의 경우, deep prompt tuning이 BERT-base에서는 미세 조정에 비해 훨씬 열세를 보였지만, RoBERTa-base에서는 1.5~2% 가량 낮지만 유사한 성능을 보였고 large에서는 0.3~0.5% 가량 미세 조정을 앞서는 모습을 보였다. 본 결과를 통해 사전학습 모델이 학습한 정보가 많을수록, 기존의 학습한 내용을 동결시킨 상태로 새롭게 프롬프트를 학습하여 추론하는 deep prompt tuning이 성능 우위를 갖는 것을 확인할 수 있다.

4.4 결과 분석

4.4.1 Prompt 길이

Deep prompt tuning에서 prompt의 길이가 너무 짧으면 전이학습이 잘 이루어지지 않고 너무 길면 입력의 최대 길이를 넘거나 학습에 악영향을 줄 수 있기 때문에, prompt의 길이는 학습에서 중요한 요인이다. 본 연구와

그림 2 Prompt 길이에 따른 성능 변화



[1] DPT의 파라미터 개수는 각 모델에 가장 좋은 성능의 파라미터 길이를 적용했을 때의 개수이다. 그림 2 참조

문서	질의	FT		DPT	
		답변	확률	답변	확률
6269760-32-0	1994년 미국에 의해 추진된 대북 폭격 논의는 어느 지역을 향한 것이었나?	"**영변**",	0.990	"영변 원자로 시설"	0.231
		"영변 원자로 시설"	0.008	"**영변**"	0.222
6584295-16-0	한국의 독립운동가 박열의 부인은?	"가네코 후미코와 한국의 독립운동가 박열"	0.270	"**가네코 후미코**" (0.870)	0.870
		"가네코 후미코"	0.257	"아나키스트 가네코 후미코"	0.094
6526255-2-1	1950년 월드컵 때 결승전에서 우루과이와 경기를 한 나라는?	"스페인"	0.979	"스웨덴"	0.191
		"**브라질**"	0.007	"**브라질**"	0.105

표 3. 미세 조정과 Deep prompt tuning의 오류
(굵은 글씨: 실제 정답 답변, 밑줄: 모델의 예측)

유사한 실험을 진행한 [8]의 경우, 추출형 질의응답을 위한 prompt 길이를 SQuAD 1.0[11]에서 18을, SQuAD 2.0[12]에서 8을 활용한 것을 확인하였다. 이를 참고하여 본 연구는 prompt 길이의 범위를 1~20으로 설정하고 전수조사를 시행하였다. 결과는 그림 2와 같다.

KLUE/BERT-base의 경우, 전반적인 성능이 prompt 길이에 비례하여 상향하였으며 F_1 점수와 완전 일치 비율은 각각 prompt의 길이가 17일 때 높은 성능을 보였다. KLUE/RoBERTa-base의 경우, prompt 길이가 14인 경우 양 평가 지표 모두에서 특별히 가장 높은 성능을 보였다. KLUE/RoBERTa-large의 경우 base 크기에 비해 상대적으로 prompt 길이에 영향을 덜 받았으며, prompt 길이가 14인 경우가 가장 높은 성능을 보였다.

4.4.2 오류 분석

Deep prompt tuning이 어떠한 차이로 미세 조정과 비교하여 성능이 올랐는지 조사하기 위해 가장 성능이 좋은 KLUE/RoBERTa-large 모델을 대상으로 최종 예측 확률이 낮은 경우, 즉 모델이 낮은 확신으로 답변한 예시를 유형별로 나누어 선택하고 그 오류 결과을 정성적으로 분석하였다. 답변의 예측 확률을 낮은 순으로 정렬하였는데, 미세 조정을 진행한 조건에서 가장 낮은 20개, deep prompt tuning을 진행한 조건에서 가장 낮은 20개, 두 조건 모두에서 가장 낮은 경우의 20개, 합하여 총 60개를 선택하였다. 표 3는 이러한 3가지 유형의 대표적인 예시 한 개씩 나열한 것으로, 실제 정답과 모델이 예측한 정답, 그에 대한 확률을 나타낸다. Deep prompt tuning은 가장 높은 확률의 답변이 오답인 경우에, 그 답변이 실제 정답과 유사하거나 그 다음으로 높은 답변 후보지가 실제 정답인 경우가 많았다. (표 3.의 문서 6269760-32-0, 6584295-16-0) Deep prompt tuning과 미세 조정이 모두 틀린 고난이도 질문의 경우에도, 실제 정답 답변에 대한 확률은 deep prompt tuning이 대부분의 경우 높게 나타났다. (표 3.의 문서 6526255-2-1)

종합적으로, deep prompt tuning이 상대적으로 난이도가 높은 질문에 대해 더욱 신중하게 답변한다는 사실을 확인하였다. 따라서 최종적인 예측 한 개 만을 고려하는 평가 지표가 아닌, 모델의 여러 예측 중 실제 정답의 순위를 평가하는 평균 상호 순위(Mean reciprocal Rank)등으로 평가할 경우 deep prompt tuning이 미세 조정보다 더욱 큰 차이의 성능 우위를 보일 수 있다. 또한 답변 난이도가 높아 미세 조정과 deep prompt tuning 모두에서 오답이 답변으로 도출되는 경우, 후처리를 비롯한 방법을 통해 실제 정답을 예측하도록 수정할 수 있는 가능성이 deep prompt tuning이 미세 조정보다 훨씬 높다.

5. 결론

본 논문에서는 기존의 미세 조정 대신 deep prompt tuning이라는 전이학습을 통해 한국어 추출형 질의응답 작업을 수행하였다. 결과적으로 학습 파라미터의 개수를 0.2~0.3%로 현저히 줄이고 학습시간을 20~30% 단축시킴과 동시에 large 크기의 모델에 대해서는 성능 향상을 이룰 수 있었다. 본 연구를 통해 deep prompt tuning 모델링이 한국어 기계 독해 작업에 적용될 가능성을 확인할 수 있었다. 향후 연구에서는 이를 다중 작업 학습(multi task learning)에 적용하기 위한 방법을 연구하고자 한다.

감사의 글

이 성과는 2023 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF- 2022R1A2C1005316)

참고 문헌

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 30. 2017.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", JMLR: Journal of Machine Learning Research, Vol. 21, No. 140, pp.1-61, 2020.

[3] Howard, J., & Ruder, S. Universal Language Model Fine-tuning for Text Classification. Universal Language Model Fine-tuning for Text Classification. https://doi.org/10.18653/v1/p18-1031, 2018

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Ka- plan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learn- ers," Advances in neural information processing sys- tems, Vol. 33, pp. 1877-1901, 2020.

[5] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for Parameter-Efficient Prompt Tuning," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, doi: 10.18653/v1/2021.emnlp-main.243, 2021

[6] G. Qin and J. Eisner, "Learning How to Ask: Querying LMs with Mixtures of Soft Prompts," Learning How to Ask: Querying LMs With Mixtures of Soft Prompts, doi: 10.18653/v1/2021.naacl-main.410. 2021

[7] Li, X. L., & Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190. 2021

[8] LIU, Xiao, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-Tuning Universally Across Scales and Tasks. *arXiv:2110.07602*. 2021

[9] S. Lim, M. Kim, and J. Lee, "KORQUAD1.0: Korean QA Dataset for machine Reading Comprehension.," arXiv (Cornell University), Sep. 2019

[10] K. Lee, S. Salant, T. Kwiatkowski, A. P. Parikh, D. Das, and J. Berant, "Learning recurrent span representations for extractive question answering," arXiv (Cornell University), Nov. 2016

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQUAD: 100,000+ questions for machine comprehension of text," arXiv (Cornell University), Jun. 2016, doi: 10.48550/arxiv.1606.05250.

[12] Pranav Rajpurkar, Robin Jia, and Percy Liang,. "Know What You Don't Know: Unanswerable Questions for SQuAD" In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784-789, Melbourne, Australia. Association for Computational Linguistics. 2018

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186, Jun. 2019.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," CoRR, Vol. abs/1907.11692, 2019.

[15] SHIN, Talyor, et al. AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *arXiv:2010.15980*

[16] Chen, Y. (n.d.). Exploring lottery prompts for pre-trained language models - arxiv.org. https://arxiv.org/pdf/2305.19500.pdf, 2023

[17] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J. Ha, and K. Cho, "Klue: Korean lan- guage understanding evaluation," 2021.