DACON

DATA TO VALUE

인하 챌린지

한국 경제 기사 분석 및 질의응답

Explore Now

DACON
DATA TO VALUE

목차

과정 요약

최종 결과물 접근 과정

01 LLM 기본 모델 선정

02 파인튜닝

03 모델 별 추론

04 앙상블 기법

LLM 기본 모델 비교

f1 score

Nous

EEVE

kosolar

gemma2

llama2

model Tuning

01

튜닝할 모델 선정

02

학습용 프롬프트 조정

03

파인튜닝 진행

model Tuning

학습용 프롬프트 조정

Gemma2 모델

<bos><start_of_turn>user
너는 주어진 Context를 토대로 Question에 답하는 챗봇이야. Question에 대한 답변만 가급적 한 단어로 최대한 간결하게 답변하도록 해.:

{본문}

Question:
{질문}<end_of_turn>
<start_of_turn>model
{답변}<end_of_turn><eos>

model Tuning

학습용 프롬프트 조정

EEVE 모델

<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
너는 주어진 Context를 토대로 Question에 답하는 챗봇이야. Question에 대한 답변만 가급적 한 단어로 최대한 간결하게 답변하도록 해.:

{본문}

Question:
{질문}<|im_end|>
<|im_start|>assistant
{답변}<|im_end|><eos>

model Tuning

학습용 프롬프트 조정

Nous 모델

“<s>### System:₩n{본문}₩n₩n### User:₩n{질문}₩n### Assistant:₩n{답변}</s>”

model Tuning

4비트 양자화 QLoRA 설정

```
lora_config = LoraConfig(
    r=6,
    target_modules=["q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj"],
    task_type="CAUSAL_LM",
)

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.float16
)
```

```
BASE_MODEL = "          모델명          "
model = AutoModelForCausalLM.from_pretrained(BASE_MODEL, device_map="auto", quantization_config=bnb_config)
tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL)
tokenizer.padding_side = 'right'
```

model Tuning

Trainer 실행

```
trainer = SFTTrainer(
    model=model,
    train_dataset=train_data,
    max_seq_length=None,
    args=TrainingArguments(
        output_dir="gemma2_results",
        # 빠른 테스트를 위해 epoch 값 1로 고정
        num_train_epochs = 1,
        max_steps=-1,
        per_device_train_batch_size=1,
        gradient_accumulation_steps=1,
        optim="paged_adamw_32bit",
        warmup_ratio=0.03,
        learning_rate=2e-4,
        fp16=True,
        logging_steps=1000,
        push_to_hub=False,
        report_to='none',
    ),
    peft_config=lora_config,
    formatting_func=generate_prompt,
)
```

```
trainer.train()
```

Inference

추론용 프롬프트

Gemma2 모델, EEVE 모델

"너는 주어진 Context를 토대로 Question에 답하는 챗봇이야. Question에 대한 답변만 가급적 한 단어로 최대한 간결하게 답변하도록 해.:₩n₩n{본문} ₩n₩nQuestion:₩n{질문}"

nous 모델

"너는 주어진 Context를 토대로 Question에 답하는 챗봇이야. ₩
Question에 대한 답변만 최대한 한 단어로, 명사들로만 답변해줘. ₩
예를 들면 누구(사람에 대한 질문)-> 김철수(사람 이름만, 직무 회사 제외),
물건 -> 사과, 시간 -> 1시, 속력 -> 1m/s, 이런식으로 딱 간단하게만 나타내줘. ₩
System:₩n{본문}₩n₩n### User:₩n{질문}₩n₩n"

LLM 기본 모델 비교

f1 score

Nous

EEVE

kosolar

gemma2

llama2

gemma2

Nous

EEVE

kosolar

파인튜닝 후 모델 비교

f1 score

gemma2

Nous

EEVE

kosolar

파인튜닝 후 모델 비교

f1 score

Ensemble model

01 ko-gemma-2-9b-it

02 Nous-Hermes-2-SOLAR-10.7B

03 yanolja EEVE-Korean-Instruct-10.8B-v1.0

progress 01

Stacking

progress 02

Voting
동점 시 젬마2 결과
단순 선택

progress 03

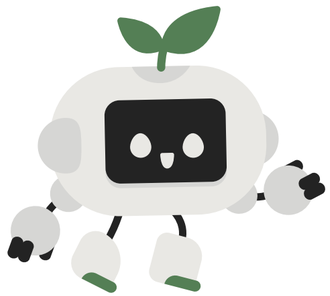
Voting
동점 시 f1 score
합이 큰 답 선택

Ensemble

Reason

단일 모델 성능 한계
-> Ensemble 기법을
통해 상호 보완

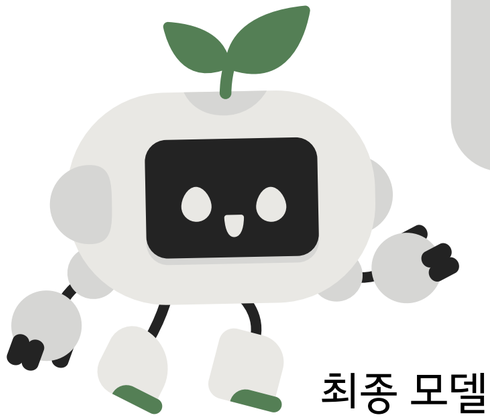
+DATA



고양이는
귀엽고 사랑스러워

+DATA

고양이는 귀엽고 사랑스럽고
부들부들해



최종 모델

Stacking

Ensemble

progress 01

Stacking

progress 02

Voting

동점 시 젬마2 결과
단순 선택

progress 03

Voting

동점 시 f1 score
합이 큰 답 선택

Reason

단일 모델 성능 한계
-> Ensemble 기법을
통해 상호 보완

DACON
DATA TO VALUE

Voting case

1087	3 : 0 : 0
322	2 : 1 : 0
98	1 : 1 : 1

1507

DACON
DATA TO VALUE

f1 score

```python
def f1_score(prediction, ground_truth):
    prediction_tokens = normalize_answer(prediction).split()
    ground_truth_tokens = normalize_answer(ground_truth).split()

    prediction_Char = []
    for tok in prediction_tokens:
        now = [a for a in tok]
        prediction_Char.extend(now)

    ground_truth_Char = []
    for tok in ground_truth_tokens:
        now = [a for a in tok]
        ground_truth_Char.extend(now)

    common = Counter(prediction_Char) & Counter(ground_truth_Char)
    num_same = sum(common.values())
    if num_same == 0:
        return 0

    precision = 1.0 * num_same / len(prediction_Char)
    recall = 1.0 * num_same / len(ground_truth_Char)
    f1 = (2 * precision * recall) / (precision + recall)

    return f1
```

DACON
DATA TO VALUE

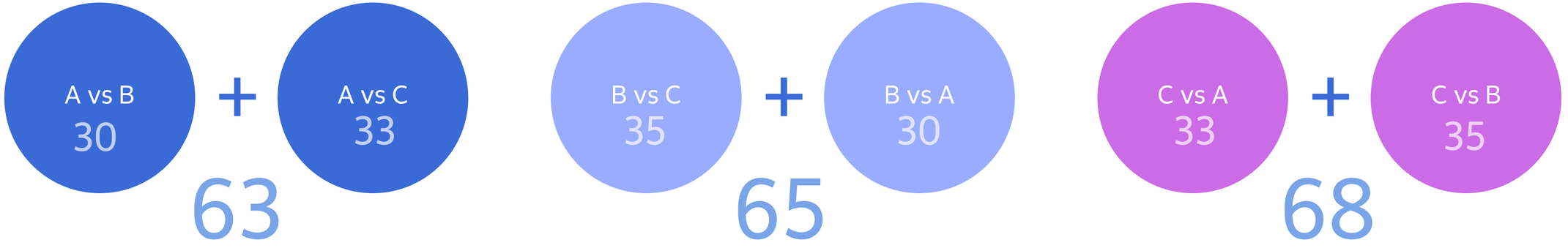
Ensemble
최종 답안 과정

01

2개 모델의 f1 score를 계산

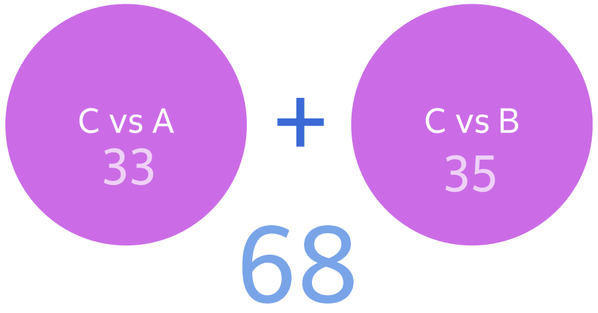
02

모델별 f1 score의 합을 계산



03

f1 score의 합이 가장 높은 답안
-> 최종 답안



Ensemble EX

id	모델A	모델B	모델C	A vs B	B vs C	A vs C	최종 정답
TEST_0001	중국과 일본	중국과 일본	중국과 일본	1	1	1	중국과 일본
TEST_0008	25명	25명	1710명	1	0.25	0.25	25명
TEST_0020	다섯 차례	14일	14일, 다섯 차례	0	0.6	0.7272727273	14일, 다섯 차례
TEST_0099	공소시효	선거사범의 공소시효	선거 수사의 연속성	0.6153846154	0.4705882353	0	선거사범의 공소시효
TEST_0156	미국 경제 침체도 그 책임을 중국에 전가하게 될 것	코로나19로 인한 미국 경제 침체	노골적인 보호무역주의와 양자주의, 중국에 대한 제재와 미·중 무역갈등은 더욱 심해질 것	0.3529411765	0.1960784314	0.2105263158	미국 경제 침체도 그 책임을 중국에 전가하게 될 것

최종 점수

ensemble

실패한 시도

01

형태소 분석기를 사용한 데이터 전처리 기법

02

QA모델 사용

03

내부 데이터를 이용한 R.A.G 진행

Q&A

감사합니다