허깅페이스 과제

2020147036 정석호

1. CV모델 불러오기 + 실행

첫 번째 모델 코드

```
from PIL import Image
from transformers import BlipProcessor, BlipForConditionalGeneration

# 모델과 프로세서 로드
processor = BlipProcessor.from_pretrained("Salesforce/blip-image-captioning-large")
model =
BlipForConditionalGeneration.from_pretrained("Salesforce/blip-image-captioning-large")

# 로컬 파일 시스템에서 이미지 불러오기
img_path = "C:/Users/user/Desktop/KakaoTalk_20240111_123152411.jpg"
raw_image = Image.open(img_path).convert('RGB')

# 조건부 이미지 캡셔닝
text = "a photography of"
inputs = processor(raw_image, text, return_tensors="pt")

out = model.generate(**inputs)
print(processor.decode(out[0], skip_special_tokens=True))

# 비조건부 이미지 캡셔닝
inputs = processor(raw_image, return_tensors="pt")

out = model.generate(**inputs)
print(processor.decode(out[0], skip_special_tokens=True))
```

첫 번째 모델 실행 결과

C:\Users\user\PycharmProjects\pythonProject4\venv\Scripts\python.exe C:\Users\user\PycharmProjects\pythonProject4\main.py

config.json: 100%|████████████████| 4.60k/4.60k [00:00<00:00, 4.52MB/s]

C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\huggingface_hub\file_download.py:149: UserWarning: `huggingface_hub` cache-system uses symlinks by default to efficiently store duplicated files but your machine does not support them in C:\Users\user\.cache\huggingface\hub\models--Salesforce--blip-image-captioning-large. Caching files will still work but in a degraded version that might require more space on your disk. This warning can be disabled by setting the `HF_HUB_DISABLE_SYMLINKS_WARNING` environment variable. For more details, see https://huggingface.co/docs/huggingface_hub/how-to-cache#limitations.

To support symlinks on Windows, you either need to activate Developer Mode or to run Python as an administrator. In order to see activate developer mode, see this article:

```
https://docs.microsoft.com/en-us/windows/apps/get-started/enable-your-device-for-development
  warnings.warn(message)
model.safetensors: 100%|██████████████| 1.88G/1.88G [06:37<00:00, 4.72MB/s]
C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\transformers\generation\utils.py:1355: UserWarning: Using the model-agnostic default `max_length` (=20) to control the generation length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
a photography of a person petting a small animal on a blanket
C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\transformers\generation\utils.py:1355: UserWarning: Using the model-agnostic default `max_length` (=20) to control the generation length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
someone is holding a baby raccoon in a blanket on a floor

Process finished with exit code 0
```

두 번째 모델 코드

```python
from PIL import Image
from transformers import BlipProcessor, BlipForConditionalGeneration

# 모델과 프로세서 로드
processor = BlipProcessor.from_pretrained("Salesforce/blip-image-captioning-base")
model =
BlipForConditionalGeneration.from_pretrained("Salesforce/blip-image-captioning-base")

# 로컬 파일 시스템에서 이미지 불러오기
img_path = "C:\\Users\\user\\Desktop\\KakaoTalk_20240111_123152411.jpg"
raw_image = Image.open(img_path).convert('RGB')

# 조건부 이미지 캡셔닝
text = "a photography of"
inputs = processor(raw_image, text, return_tensors="pt")

out = model.generate(**inputs)
print(processor.decode(out[0], skip_special_tokens=True))

# 비조건부 이미지 캡셔닝
inputs = processor(raw_image, return_tensors="pt")

out = model.generate(**inputs)
print(processor.decode(out[0], skip_special_tokens=True))
```

두 번째 모델 코드 실행 결과

```
C:\Users\user\PycharmProjects\pythonProject4\venv\Scripts\python.exe C:\Users\user\PycharmProjects\pythonProject4\main.py
preprocessor_config.json: 100%|██████████| 287/287 [00:00<?, ?B/s]
C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\huggingface_hub\file_download.py:149: UserWarning: `huggingface_hub` cache-system uses symlinks by default to efficiently store duplicated files but your machine does not support them in C:\Users\user\.cache\huggingface\hub\models--Salesforce--blip-image-captioning-base. Caching files will still work but in a degraded version that might require more space on your disk. This warning can be disabled by setting the `HF_HUB_DISABLE_SYMLINKS_WARNING` environment variable. For more details, see https://huggingface.co/docs/huggingface_hub/how-to-cache#limitations.
To support symlinks on Windows, you either need to activate Developer Mode or to run Python as an administrator. In order to see activate developer mode, see this article: https://docs.microsoft.com/en-us/windows/apps/get-started/enable-your-device-for-development
  warnings.warn(message)
tokenizer_config.json: 100%|██████████| 506/506 [00:00<?, ?B/s]
vocab.txt: 100%|██████████| 232k/232k [00:00<00:00, 3.50MB/s]
tokenizer.json: 100%|██████████| 711k/711k [00:00<00:00, 2.73MB/s]
special_tokens_map.json: 100%|██████████| 125/125 [00:00<00:00, 125kB/s]
config.json: 100%|██████████| 4.56k/4.56k [00:00<00:00, 4.60MB/s]
pytorch_model.bin: 100%|██████████| 990M/990M [03:55<00:00, 4.20MB/s]
C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\transformers\generation\utils.py:1355: UserWarning: Using the model-agnostic default `max_length` (=20) to control the generation length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
a photography of a baby raccoon
C:\Users\user\PycharmProjects\pythonProject4\venv\Lib\site-packages\transformers\generation\utils.py:1355: UserWarning: Using the model-agnostic default `max_length` (=20) to control the generation length. We recommend setting `max_new_tokens` to control the maximum length of the generation.
  warnings.warn(
a person holding a baby racco

Process finished with exit code 0
```

모델 성능 비교 :
같은 사진을 보고 첫 번째 모델은 ‘someone is holding a baby raccoon in a blanket on a floor’이라는 분석을, 두 번째 모델은 ‘a person holding a baby raccoon’와 같은 분석을 했다.
Hugging Face 사이트에서 large 모델과 base 모델을 가져왔는데 확실히 large모델이 더 상세한 분석을 해냈다.

코드에 활용한 사진 :

2. NLP 모델 FINE-TUNING

데이터전처리 과정에서 막혔습니다...

[전처리 반영 전 코드]

```
from transformers import pipeline
unmasker = pipeline('fill-mask', model='bert-base-uncased')
unmasker("Hello I'm a [MASK] model.")

from transformers import BertTokenizer, BertModel
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained("bert-base-uncased")
text = "Although it almost entirely follows actual historical events, which can be infuriating to watch, it's artistically intriguing and highly immersive due to its excellent composition. Every time I watch it, I find something new and interesting. It's a work that captures both meaning and entertainment!"
encoded_input = tokenizer(text, return_tensors='pt')
output = model(**encoded_input)
```

[전처리할 데이터 출처]

https://huggingface.co/datasets/ajaykarthick/imdb-movie-reviews

[전처리 코드]

```
from transformers import BertTokenizer
import os

# BERT 토크나이저 초기화
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

# 텍스트 파일이 있는 디렉토리(여기에 디렉토리 경로를 입력하세요)
directory_path = "C:/Users/user/Desktop/neg"

# 디렉토리 내의 모든.txt 파일 찾기
file_paths = [os.path.join(directory_path, file) for file in os.listdir(directory_path) if file.endswith('.txt')]

# 모든 파일에서 텍스트 읽기
texts = []
for file_path in file_paths:
with open(file_path, 'r', encoding='utf-8') as file:
texts.extend([line.strip() for line in file])

# 토큰화 및 특수 토큰 추가, 패딩, 길이 정규화
inputs = tokenizer(texts, padding=True, truncation=True, return_tensors="pt")

# 입력 데이터
print(inputs)
```