ng2fim55b

November 27, 2023

0.1 # Entrenamiento de un transformer para Q&A

- Diego Sú Gómez - A01620476
- Estefanía Pérez Yeo - A01639270
- Isai Ambrocio - A01625101
- Vanessa Méndez Palacios - A01639925
- Francisco Javier Sanchez Panduro - A01639832

[ ]:
```
!pip install datasets transformers transformers[torch]
```

0.1.1 Importación de librerías

[2]:
```
# Inicio de sesión con Hugging Face para cargar el modelo
from huggingface_hub import notebook_login

notebook_login()
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/
↪assets/huggingface_logo-noborder.sv…
```

[3]:
```
import transformers
import torch
from datasets import load_dataset, load_metric
from transformers import BertForQuestionAnswering, AutoTokenizer,␣
↪default_data_collator,  TrainingArguments, Trainer
from torch.utils.data import DataLoader, Dataset
```

0.1.2 Obtención de corpus

[ ]:
```
dataset = load_dataset("squad_v2")
```

[5]:
```
dataset
```

[5]:
```
DatasetDict({
    train: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 130319
    })
```

```
    validation: Dataset({
        features: ['id', 'title', 'context', 'question', 'answers'],
        num_rows: 11873
    })
})
```

[6]:
```
# Carga de un elemento del dataset de entrenamiento
dataset["train"][0]
```

[6]:
```
{'id': '56be85543aeaaa14008c9063',
 'title': 'Beyoncé',
 'context': 'Beyoncé Giselle Knowles-Carter (/bi j nse / bee-YON-say) (born
September 4, 1981) is an American singer, songwriter, record producer and
actress. Born and raised in Houston, Texas, she performed in various singing and
dancing competitions as a child, and rose to fame in the late 1990s as lead
singer of R&B girl-group Destiny\'s Child. Managed by her father, Mathew
Knowles, the group became one of the world\'s best-selling girl groups of all
time. Their hiatus saw the release of Beyoncé\'s debut album, Dangerously in
Love (2003), which established her as a solo artist worldwide, earned five
Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in
Love" and "Baby Boy".',
 'question': 'When did Beyonce start becoming popular?',
 'answers': {'text': ['in the late 1990s'], 'answer_start': [269]}}
```

0.1.3 Preprocesamiento del dataset

[7]:
```
# Generación del tokenizer para dar formato a los datos
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")
```

```
tokenizer_config.json:   0%|          | 0.00/28.0 [00:00<?, ?B/s]
config.json:   0%|          | 0.00/570 [00:00<?, ?B/s]
vocab.txt:   0%|          | 0.00/232k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/466k [00:00<?, ?B/s]
```

[8]:
```
# Función para procesar todos los datos del dataset y aplicarlos en el modelo
def prepare_train_features(data):
  pad = tokenizer.padding_side == "right"
  data["question"] = [q.lstrip() for q in data["question"]]

  lang = "es" if "¿" in data["question"][0] else "en"

  if lang == "en":
    tokenizer_lang = AutoTokenizer.from_pretrained("bert-base-uncased")
  elif lang == "es":
```

```
  tokenizer_lang = AutoTokenizer.from_pretrained("dbmdz/
↪bert-large-spanish-wwm-cased")
else:
  raise ValueError("Idioma no compatible")

tokenized_data = tokenizer_lang(
  data["question" if pad else "context"],
  data["context" if pad else "question"],
  truncation="only_second" if pad else "only_first",
  max_length=385,
  stride=128,
  return_overflowing_tokens=True,
  return_offsets_mapping=True,
  padding="max_length",
  )

sample_mapping = tokenized_data.pop("overflow_to_sample_mapping")
offset_mapping = tokenized_data.pop("offset_mapping")

tokenized_data["start_positions"] = []
tokenized_data["end_positions"] = []

for i, offsets in enumerate(offset_mapping):
    input_ids = tokenized_data["input_ids"][i]
    cls_index = input_ids.index(tokenizer.cls_token_id)

    sequence_ids = tokenized_data.sequence_ids(i)

    sample_index = sample_mapping[i]
    answers = data["answers"][sample_index]
    if len(answers["answer_start"]) == 0:
        tokenized_data["start_positions"].append(cls_index)
        tokenized_data["end_positions"].append(cls_index)
    else:
        start_char = answers["answer_start"][0]
        end_char = start_char + len(answers["text"][0])

        token_start_index = 0
        while sequence_ids[token_start_index] != (1 if pad else 0):
            token_start_index += 1

        token_end_index = len(input_ids) - 1
        while sequence_ids[token_end_index] != (1 if pad else 0):
            token_end_index -= 1

        if not (offsets[token_start_index][0] <= start_char and␣
↪offsets[token_end_index][1] >= end_char):
```

```
            tokenized_data["start_positions"].append(cls_index)
            tokenized_data["end_positions"].append(cls_index)
        else:
            while token_start_index < len(offsets) and␣
↪offsets[token_start_index][0] <= start_char:
                token_start_index += 1
            tokenized_data["start_positions"].append(token_start_index - 1)
            while offsets[token_end_index][1] >= end_char:
                token_end_index -= 1
            tokenized_data["end_positions"].append(token_end_index + 1)

    return tokenized_data
```

[9]:
```
tokenized_dataset = dataset.map(prepare_train_features, batched = True,␣
↪remove_columns = dataset["train"].column_names)
```

```
Map:   0%|          | 0/130319 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/11873 [00:00<?, ? examples/s]
```

0.1.4 Fine-Tuning BERT

[10]:
```
model = BertForQuestionAnswering.from_pretrained("bert-base-uncased")
```

```
model.safetensors:   0%|          | 0.00/440M [00:00<?, ?B/s]
```

```
Some weights of BertForQuestionAnswering were not initialized from the model
checkpoint at bert-base-uncased and are newly initialized: ['qa_outputs.bias',
'qa_outputs.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

[11]:
```
# Definición de argumentos para el entrenamiento
args = TrainingArguments("BERT-Finetuned", evaluation_strategy="epoch",␣
↪learning_rate=2e-5, per_device_train_batch_size=16, num_train_epochs=3,␣
↪weight_decay=0.05)
```

[12]:
```
# Generación de batches
data_collator = default_data_collator
```

[13]:
```
# División del dataset de entrenamiento para reducir el tiempo de ejecución
tokenized_train = tokenized_dataset["train"].select(list(range(10000)))
```

[14]:
```
# División del dataset de validación para reducir el tiempo de ejecución
tokenized_val = tokenized_dataset["validation"].select(list(range(10000)))
```

[15]:
```
# Definición del entrenador del modelo
```

```
trainer = Trainer(model, args, train_dataset = tokenized_train,␣
 ↪eval_dataset=tokenized_val, data_collator=data_collator, tokenizer=tokenizer)
```

[16]:
```
trainer.train()
```

```
<IPython.core.display.HTML object>
```

[16]:
```
TrainOutput(global_step=1875, training_loss=1.37319267578125,
metrics={'train_runtime': 2996.8506, 'train_samples_per_second': 10.011,
'train_steps_per_second': 0.626, 'total_flos': 5894487383400000.0, 'train_loss':
1.37319267578125, 'epoch': 3.0})
```

[17]:
```
model.save_pretrained("BERT-FinalModel")
```

0.1.5 Preguntas

[18]:
```
def answer_questions(question, context, q_no):
  inputs = tokenizer(question, context, return_tensors='pt')
  inputs = {key: value.to(model.device) for key, value in inputs.items()}

  outputs = model(**inputs)
  start_scores = outputs.start_logits
  end_scores = outputs.end_logits

  start_index = torch.argmax(start_scores)
  end_index = torch.argmax(end_scores)

  tokens = inputs['input_ids'][0][start_index:end_index + 1]
  answer_tokens = tokenizer.convert_ids_to_tokens(tokens.tolist())
  answer = tokenizer.convert_tokens_to_string(answer_tokens)

  print(f"Question {q_no + 1}: {question}")
  print(f"Answer{q_no + 1}: {answer}")
  print("\n")
```

Inglés

[19]:
```
questions_analysis_en = [
    "What is the primary goal of artificial intelligence?",
    "Quick left ventricle analysis in a clinical setting provides timely␣
 ↪information for early diagnosis, allowing prompt intervention and improved␣
 ↪patient care.",
    "Explain the process of creating masks in the analysis of left ventricles.",
    "What are the important landmarks in the analysis of left ventricles?",
    "How does machine learning contribute to artificial intelligence?",
    "What information can be obtained from the analysis of left ventricles?",
```

```
    "What is the clinical impact of analyzing left ventricles with artificial␣
 ↪intelligence?",
    "Why are landmarks important in medical imaging, especially in cardiac␣
 ↪analysis?",
    "What are the common challenges in the analysis of left ventricles?",
    "What is the primary purpose of analyzing the left ventricle?",
]

contexts_analysis_en = [
    "The primary goal of artificial intelligence is to develop systems that can␣
 ↪perform tasks that typically require human intelligence, such as␣
 ↪problem-solving, learning, and decision-making.",
    "The analysis of left ventricles with artificial intelligence involves the␣
 ↪use of advanced algorithms to process data and extract information.",
    "Creating masks in the analysis of left ventricles is a crucial step to␣
 ↪identify regions of interest and improve analysis accuracy.",
    "In the analysis of left ventricles, landmarks are key anatomical points␣
 ↪used for precise evaluation and measurement.",
    "Machine learning is a subset of artificial intelligence that focuses on␣
 ↪developing algorithms and models that allow systems to learn from data,␣
 ↪improving their performance over time.",
    "From the analysis of left ventricles, information about contractile␣
 ↪function, volume, and overall heart health can be obtained.",
    "The clinical impact of analyzing left ventricles is significant, aiding in␣
 ↪the early diagnosis of heart diseases.",
    "In medical imaging, particularly in cardiac analysis, landmarks play a␣
 ↪crucial role. These specific points serve as reference markers, aiding in␣
 ↪accurate measurements, precise evaluations, and facilitating consistent␣
 ↪analysis across different cases.",
    "Common challenges in the analysis of left ventricles include anatomical␣
 ↪variability and image quality.",
    "The primary purpose of analyzing the left ventricle is to assess its␣
 ↪function and structure, providing insights into overall heart health.",
    ]
```

[20]:
```
for q_no in range(10):
  answer_questions(questions_analysis_en[q_no], contexts_analysis_en[q_no],␣
 ↪q_no)
```

```
Question 1: What is the primary goal of artificial intelligence?
Answer1: to develop systems that can perform tasks that typically require human
intelligence, such as problem - solving, learning, and decision - making


Question 2: Quick left ventricle analysis in a clinical setting provides timely
information for early diagnosis, allowing prompt intervention and improved
```

```
patient care.
Answer2: [CLS]


Question 3: Explain the process of creating masks in the analysis of left
ventricles.
Answer3: [CLS]


Question 4: What are the important landmarks in the analysis of left ventricles?
Answer4: anatomical points


Question 5: How does machine learning contribute to artificial intelligence?
Answer5: developing algorithms and models


Question 6: What information can be obtained from the analysis of left
ventricles?
Answer6: contractile function, volume, and overall heart health


Question 7: What is the clinical impact of analyzing left ventricles with
artificial intelligence?
Answer7: significant


Question 8: Why are landmarks important in medical imaging, especially in
cardiac analysis?
Answer8: a crucial role


Question 9: What are the common challenges in the analysis of left ventricles?
Answer9: anatomical variability and image quality


Question 10: What is the primary purpose of analyzing the left ventricle?
Answer10: to assess its function and structure
```

Español

[21]:
```
questions_analysis_spa = [
    "¿Cuál es el objetivo principal de la inteligencia artificial?",
    "El análisis rápido del ventrículo izquierdo en un entorno clínico␣
 ↪proporciona información oportuna para un diagnóstico temprano, permitiendo␣
 ↪una intervención rápida y una mejor atención al paciente.",
```

```
    "Explica el proceso de creación de máscaras en el análisis de ventrículos␣
 ↪izquierdos.",
    "¿Cuáles son los puntos de referencia importantes en el análisis de␣
 ↪ventrículos izquierdos?",
    "¿Cómo contribuye el aprendizaje automático a la inteligencia artificial?",
    "¿Qué información se puede obtener del análisis de ventrículos izquierdos?",
    "¿Cuál es el impacto clínico de analizar los ventrículos izquierdos con␣
 ↪inteligencia artificial?",
    "¿Por qué son importantes los puntos de referencia en la imagen médica,␣
 ↪especialmente en el análisis cardíaco?",
    "¿Cuáles son los desafíos comunes en el análisis de ventrículos izquierdos?
 ↪",
    "¿Cuál es el propósito principal de analizar el ventrículo izquierdo?",
]

contexts_analysis_spa = [
    "El objetivo principal de la inteligencia artificial es desarrollar␣
 ↪sistemas que puedan realizar tareas que típicamente requieren inteligencia␣
 ↪humana, como resolver problemas, aprender y tomar decisiones.",
    "El análisis de ventrículos izquierdos con inteligencia artificial implica␣
 ↪el uso de algoritmos avanzados para procesar datos y extraer información.",
    "Crear máscaras en el análisis de ventrículos izquierdos es un paso crucial␣
 ↪para identificar regiones de interés y mejorar la precisión del análisis.",
    "En el análisis de ventrículos izquierdos, los puntos de referencia son␣
 ↪puntos anatómicos clave utilizados para una evaluación y medición precisa.",
    "El aprendizaje automático es un subconjunto de la inteligencia artificial␣
 ↪que se centra en desarrollar algoritmos y modelos que permitan a los␣
 ↪sistemas aprender de los datos, mejorando su rendimiento con el tiempo.",
    "Del análisis de ventrículos izquierdos se puede obtener información sobre␣
 ↪la función contráctil, el volumen y la salud general del corazón.",
    "El impacto clínico de analizar ventrículos izquierdos es significativo,␣
 ↪ayudando en el diagnóstico temprano de enfermedades cardíacas.",
    "En la imagen médica, especialmente en el análisis cardíaco, los puntos de␣
 ↪referencia juegan un papel crucial. Estos puntos específicos sirven como␣
 ↪marcadores de referencia, ayudando en mediciones precisas, evaluaciones␣
 ↪precisas y facilitando un análisis consistente en diferentes casos.",
    "Los desafíos comunes en el análisis de ventrículos izquierdos incluyen la␣
 ↪variabilidad anatómica y la calidad de la imagen.",
    "El propósito principal de analizar el ventrículo izquierdo es evaluar su␣
 ↪función y estructura, proporcionando información sobre la salud general del␣
 ↪corazón.",
]
```

[22]:
```
for q_no in range(10):
  answer_questions(questions_analysis_spa[q_no], contexts_analysis_spa[q_no],␣
 ↪q_no)
```

Question 1: ¿Cuál es el objetivo principal de la inteligencia artificial?
Answer1: [CLS]

Question 2: El análisis rápido del ventrículo izquierdo en un entorno clínico proporciona información oportuna para un diagnóstico temprano, permitiendo una intervención rápida y una mejor atención al paciente.
Answer2: [CLS]

Question 3: Explica el proceso de creación de máscaras en el análisis de ventrículos izquierdos.
Answer3: [CLS]

Question 4: ¿Cuáles son los puntos de referencia importantes en el análisis de ventrículos izquierdos?
Answer4: [CLS]

Question 5: ¿Cómo contribuye el aprendizaje automático a la inteligencia artificial?
Answer5: [CLS]

Question 6: ¿Qué información se puede obtener del análisis de ventrículos izquierdos?
Answer6: [CLS]

Question 7: ¿Cuál es el impacto clínico de analizar los ventrículos izquierdos con inteligencia artificial?
Answer7: [CLS]

Question 8: ¿Por qué son importantes los puntos de referencia en la imagen médica, especialmente en el análisis cardíaco?
Answer8: [CLS]

Question 9: ¿Cuáles son los desafíos comunes en el análisis de ventrículos izquierdos?
Answer9: [CLS]

Question 10: ¿Cuál es el propósito principal de analizar el ventrículo izquierdo?
Answer10: [CLS]

0.1.6 Conclusiones

- **¿Hubo alguna diferencia?**

Sí, se notan varias diferencias en los resultados entre las preguntas en inglés y en español. Las respuestas en inglés son más coherentes, mientras que las preguntas en español tienen respuestas limitadas.

- **¿Qué lenguaje conviene más y por qué?**

Es más conveniente utilizar el lenguaje en inglés, ya que el modelo BERT se entrenó principalmente con datos en inglés, lo cual significa que su rendimiento será más preciso con este lenguaje.

- **¿Cuál era el tamaño del corpus?**

El corpus (SQuAD v2) tiene un tamaño de 46.49 megabytes. Sin embargo, una vez que se procesa y se genera el conjunto de datos utilizado, su tamaño total alcanza los 128.52 megabytes. Consta con un total de 142,192 filas de datos.

- **¿Cuántas respuestas tienen coherencia?**

Para las preguntas en inglés, la mayoría de las respuestas parecen ser coherentes y relacionadas con el contexto. Sin embargo, para las preguntas en español, la mayoría de las respuestas son el token [CLS]. Esto se debe a que se hicieron preguntas que abarcan un tema diferente a lo que el modelo vió durante el entrenamiento.

- **¿Si cambia el corpus y pregunta lo mismo recibirá una respuesta?**

Si optamos por cambiar el corpus y preguntar lo mismo, esto puede afectar la calidad y coherencia de las respuestas, pues la capacidad del modelo para generalizar nuevos corpus dependerá de su entrenamiento y de la similitud del nuevo corpus con los datos con los que se entrenó.

- **¿Cuántos lenguajes puede manejar el BERT para resolver preguntas?**

Como tal, no se tiene un número fijo de lenguajes que el BERT puede manejar, ya que su capacidad para manejar uno depende de cuantos modelos preentrenados hay disponibles para ese lenguaje en específico. Tomamos en cuenta que para lograr mejores resultados se debe de utilizar modelos y tokenizadores específicos para cada uno.