Entrenamiento de un transformer para Q&A

Fernando Ojeda Marín - A01639252

Mario Alberto Casteña Martínez - A01640152

Victor Hugo Arreola Elenes - A01635682

Luis Manuel Orozco Yáñez - A01707822

- Realice un reporte en equipos de reto respondiendo las siguientes preguntas.
- Obtenga una base de conocimiento o Corpus con información técnica referente a su reto.
- Utilizando un modelo previamente entrenado de “BertForQuestionAnswering”, administre como corpus el texto usado en las actividades anteriores o algún otro diferente.
- Plantee 10 preguntas que el transformer debería de responder con respecto al corpus.
- Obtenga las respuestas de esas 10 preguntas en español e inglés (recuerden que sólo se entrena una vez, la idea es ver las diferentes respuestas con entradas de diferentes idiomas):
    - ¿Hubo alguna diferencia?
    - ¿Qué lenguaje conviene más y por qué?
    - ¿Cuál era el tamaño del corpus?
    - ¿Cuántas respuestas tienen coherencia?
    - ¿Si cambia el corpus y pregunta lo mismo recibirá una respuesta? Demuestre
    - ¿Cuántos lenguajes puede manejar el BERT para resolver preguntas?

```
from transformers import BertForQuestionAnswering, BertTokenizer
import torch

# Ruta del archivo que contiene el corpus
ruta_corpus = r"D:\Tec\7mo Smestre\Inteligencia Artificial II\NLP\corpus.txt"

# Cargar el modelo preentrenado de BERT para preguntas y respuestas
modelo = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
tokenizer = BertTokenizer.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')

# Definir las preguntas en español
preguntas_espanol = [
    "¿Cuál es el mejor modelo para trabajar con imágenes?",
    "¿Cómo mejorar la generalización del modelo?",
    "¿Qué dataset se utilizó?",
    "¿Con qué tipos de datos se trabaja?",
    "¿Por qué se utiliza la CNN?",
    "¿Qué funciones de pérdida se probaron?",
    "¿Cómo se realiza la inferencia sobre el conjunto de test?",
    "¿Cuál es la métrica empleada para medir el rendimiento de los modelos?",
    "¿Cómo se puede interpretar los modelos?",
    "¿Cuáles podrían ser las conclusiones del trabajo?"
]

# Definir las preguntas en inglés
preguntas_ingles = [
    "What is the best model for working with images?",
    "How to improve the model's generalization?",
    "Which dataset was used?",
    "What types of data are being worked with?",
    "Why is CNN used?",
    "Which loss functions were tried?",
    "How is inference done on the test set?",
    "What metric is used to measure model performance?",
    "How can the models be interpreted?",
    "What could be the conclusions of the work?"
]

# Leer el contenido del corpus
```

```
with open(ruta_corpus, 'r', encoding='utf-8') as archivo:
    corpus = archivo.read()

# Tokenizar el corpus
corpus_tokens = tokenizer.encode_plus(corpus, return_tensors='pt')

# Realizar preguntas y respuestas en el corpus para preguntas en español
for pregunta in preguntas_espanol:
    # Tokenizar la pregunta
    pregunta_tokens = tokenizer.encode_plus(pregunta, return_tensors='pt')

    # Dividir el corpus en fragmentos más pequeños
    fragmentos_corpus = [corpus[i:i + 512] for i in range(0, len(corpus), 512)]

    respuestas = []

    for fragmento in fragmentos_corpus:
        # Combinar diccionarios de pregunta_tokens y fragmento del corpus
        tokens_combinados = {key: torch.cat([pregunta_tokens[key], tokenizer.encode_plus(fragmento, return_tensors='pt')[key]], dim=1) for

        # Realizar la predicción
        inicio_respuesta = modelo(**tokens_combinados)['start_logits']
        fin_respuesta = modelo(**tokens_combinados)['end_logits']

        # Decodificar los índices de inicio y fin en tokens
        inicio_respuesta_idx = inicio_respuesta.argmax()
        fin_respuesta_idx = fin_respuesta.argmax()

        # Obtener la respuesta del contexto original
        respuesta = tokenizer.decode(tokens_combinados['input_ids'][0][inicio_respuesta_idx:fin_respuesta_idx + 1])
        respuestas.append(respuesta)

    respuesta_completa = " ".join(respuestas)

    print(f"Pregunta (español): {pregunta}")
    print(f"Respuesta: {respuesta_completa}\n")

# Realizar preguntas y respuestas en el corpus para preguntas en inglés
for pregunta in preguntas_ingles:
    # Tokenizar la pregunta
    pregunta_tokens = tokenizer.encode_plus(pregunta, return_tensors='pt')

    # Dividir el corpus en fragmentos más pequeños
    fragmentos_corpus = [corpus[i:i + 512] for i in range(0, len(corpus), 512)]

    respuestas = []

    for fragmento in fragmentos_corpus:
        # Combinar diccionarios de pregunta_tokens y fragmento del corpus
        tokens_combinados = {key: torch.cat([pregunta_tokens[key], tokenizer.encode_plus(fragmento, return_tensors='pt')[key]], dim=1) for

        # Realizar la predicción
        inicio_respuesta = modelo(**tokens_combinados)['start_logits']
        fin_respuesta = modelo(**tokens_combinados)['end_logits']

        # Decodificar los índices de inicio y fin en tokens
        inicio_respuesta_idx = inicio_respuesta.argmax()
        fin_respuesta_idx = fin_respuesta.argmax()

        # Obtener la respuesta del contexto original
        respuesta = tokenizer.decode(tokens_combinados['input_ids'][0][inicio_respuesta_idx:fin_respuesta_idx + 1])
        respuestas.append(respuesta)

    respuesta_completa = " ".join(respuestas)

    print(f"Pregunta (inglés): {pregunta}")
    print(f"Respuesta: {respuesta_completa}\n")

```

```
Pregunta (español): ¿Cómo mejorar la generalización del modelo?
Respuesta: [CLS] [CLS] nto cardiaco mediante redes neuronales la deteccion de la disminucion del funcionamiento cardiaco es un factor
```

```
Pregunta (español): ¿Por qué se utiliza la CNN?
Respuesta: [CLS] un factor clave para el diagnostico de enfermedades del corazon   detection of declining heart function is a key fac

Pregunta (español): ¿Qué funciones de pérdida se probaron?
Respuesta:   analisi del funcionament cardiac [CLS] detection of declining heart function [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [

Pregunta (español): ¿Cómo se realiza la inferencia sobre el conjunto de test?
Respuesta:   ademas se han aplicado tecnicas de model interpretability para analizar el comportamiento de los modelos y poder tomar m

Pregunta (español): ¿Cuál es la métrica empleada para medir el rendimiento de los modelos?
Respuesta: [CLS] este proceso es lento y quita mucho tiempo a los medicos  [CLS]  [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CL

Pregunta (español): ¿Cómo se puede interpretar los modelos?
Respuesta:  en este trabajo se han implementado ademas se han aplicado tecnicas de model interpretability aplicat tecniques de model

Pregunta (español): ¿Cuáles podrían ser las conclusiones del trabajo?
Respuesta: [CLS] modelos basados en redes neuronales profundas [CLS]  detection of declining heart function [CLS] [CLS] [CLS]  [CLS]

Pregunta (inglés): What is the best model for working with images?
Respuesta: analisis del funcionamiento cardiaco mediante redes neuronales el volumen del ventriculo cardiac el volum del ventricle es

Pregunta (inglés): How to improve the model's generalization?
Respuesta:  el volumen del ventriculo izquierdo al final de la fase de sistole y diastole para estimar la sangre eyectada por el cora

Pregunta (inglés): Which dataset was used?
Respuesta: analisis del funcionamiento cardiaco mediante redes neuronales nto cardiaco mediante redes neuronales  xarxes neuronals pr

Pregunta (inglés): What types of data are being worked with?
Respuesta:  modelos basados en redes neuronales profundas tecnicas de model interpretability para analizar el comportamiento de los m

Pregunta (inglés): Why is CNN used?
Respuesta:  este proceso es lento y quita mucho tiempo a los medicos ademas se han aplicado tecnicas de model interpretability para a

Pregunta (inglés): Which loss functions were tried?
Respuesta: human language sistole y diastole cardiac es un factor clau per al diagnostic de malalties del cor. per a analitzar - ho s

Pregunta (inglés): How is inference done on the test set?
Respuesta: analisis del funcionamiento cardiaco mediante redes neuronales al final de la fase de sistole y diastole model interpretab

Pregunta (inglés): What metric is used to measure model performance?
Respuesta: aster volumen del ventriculo interpretability interpretability volume of the left ventricle [CLS] [CLS] [CLS] [CLS] [CLS]

Pregunta (inglés): How can the models be interpreted?
Respuesta: pattern recognition profundas para estimar el volumen del ventriculo izquierdo para analizar el comportamiento de los mode

Pregunta (inglés): What could be the conclusions of the work?
Respuesta: analisis del funcionamiento cardiaco mediante redes neuronales modelos basados en redes neuronales profundas para estimar
```

- ¿Hubo alguna diferencia? Sí, hubo diferencias en las respuestas proporcionadas. Esto puede deberse a la variabilidad en la estructura y contenido del corpus utilizado para entrenar a BertForQuestionAnswering.
- ¿Qué lenguaje conviene más y por qué? No se puede determinar con certeza cuál es el mejor lenguaje basándose únicamente en las respuestas proporcionadas. La elección del lenguaje puede depender de varios factores, como la disponibilidad y calidad del corpus en cada idioma, así como las características específicas de la tarea.
- ¿Cuál era el tamaño del corpus? 62 paginas
- ¿Cuántas respuestas tienen coherencia? Al analizar las respuestas proporcionadas a las preguntas, se observa una falta de coherencia en la mayoría de ellas. En lugar de abordar directamente los aspectos preguntados, las respuestas tienden a introducir información adicional o detalles que no son pertinentes a la pregunta original. Esto puede deberse a diversas razones, como la complejidad del modelo utilizado (BertForQuestionAnswering), la variabilidad en la formulación de las preguntas, o la naturaleza del corpus de entrenamiento.

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

Si obtenemos una respuesta sin embargo volvemos a obtener respuestas sin sentido , en este caso en mas entendible ya que las preguntas no tienen que ver con el segundo corpus pero aun asi las respuestas no tienen coherencia.

```
from transformers import BertForQuestionAnswering, BertTokenizer
import torch

# Ruta del archivo que contiene el corpus
ruta_corpus = r"D:\Tec\7mo Smestre\Inteligencia Artificial II\NLP\corpus2.txt"

# Cargar el modelo preentrenado de BERT para preguntas y respuestas
modelo = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
```

```
tokenizer = BertTokenizer.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')

# Definir las preguntas en español
preguntas_espanol = [
    "¿Cuál es el mejor modelo para trabajar con imágenes?",
    "¿Cómo mejorar la generalización del modelo?",
    "¿Qué dataset se utilizó?",
    "¿Con qué tipos de datos se trabaja?",
    "¿Por qué se utiliza la CNN?",
    "¿Qué funciones de pérdida se probaron?",
    "¿Cómo se realiza la inferencia sobre el conjunto de test?",
    "¿Cuál es la métrica empleada para medir el rendimiento de los modelos?",
    "¿Cómo se puede interpretar los modelos?",
    "¿Cuáles podrían ser las conclusiones del trabajo?"
]

# Leer el contenido del corpus
with open(ruta_corpus, 'r', encoding='utf-8') as archivo:
    corpus = archivo.read()

# Tokenizar el corpus
corpus_tokens = tokenizer.encode_plus(corpus, return_tensors='pt')

# Realizar preguntas y respuestas en el corpus para preguntas en español
for pregunta in preguntas_espanol:
    # Tokenizar la pregunta
    pregunta_tokens = tokenizer.encode_plus(pregunta, return_tensors='pt')

    # Dividir el corpus en fragmentos más pequeños
    fragmentos_corpus = [corpus[i:i + 512] for i in range(0, len(corpus), 512)]

    respuestas = []

    for fragmento in fragmentos_corpus:
        # Combinar diccionarios de pregunta_tokens y fragmento del corpus
        tokens_combinados = {key: torch.cat([pregunta_tokens[key], tokenizer.encode_plus(fragmento, return_tensors='pt')[key]], dim=1) for

        # Realizar la predicción
        inicio_respuesta = modelo(**tokens_combinados)['start_logits']
        fin_respuesta = modelo(**tokens_combinados)['end_logits']

        # Decodificar los índices de inicio y fin en tokens
        inicio_respuesta_idx = inicio_respuesta.argmax()
        fin_respuesta_idx = fin_respuesta.argmax()

        # Obtener la respuesta del contexto original
        respuesta = tokenizer.decode(tokens_combinados['input_ids'][0][inicio_respuesta_idx:fin_respuesta_idx + 1])
        respuestas.append(respuesta)

    respuesta_completa = " ".join(respuestas)

    print(f"Pregunta (español): {pregunta}")
    print(f"Respuesta: {respuesta_completa}\n")
```

```
c:\Users\vhae1\anaconda3\envs\tf\lib\site-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found. Please update jupyter and ipywidge
  from .autonotebook import tqdm as notebook_tqdm
Some weights of the model checkpoint at bert-large-uncased-whole-word-masking-finetuned-squad were not used when initializing BertForQue
- This IS expected if you are initializing BertForQuestionAnswering from the checkpoint of a model trained on another task or with anoth
- This IS NOT expected if you are initializing BertForQuestionAnswering from the checkpoint of a model that you expect to be exactly ide
Token indices sequence length is longer than the specified maximum sequence length for this model (60088 > 512). Running this sequence t
Pregunta (español): ¿Cuál es el mejor modelo para trabajar con imágenes?
Respuesta:   el principal proposito es integrar uno de estos dispositivos en un prototipo desarrollado en trabajos anteriores para obten

Pregunta (español): ¿Cómo mejorar la generalización del modelo?
Respuesta: [CLS]  [CLS]  [CLS] slam [CLS]  [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CLS] [CL

Pregunta (español): ¿Qué dataset se utilizó?
Respuesta: trabajo fin de grado titulo [CLS] light detection and ranging ( lidar simultaneous localization and mapping integrated systen

Pregunta (español): ¿Con qué tipos de datos se trabaja?
Respuesta:   las posibilidades [CLS] integrated systems in vehicles are essential for sensing the environment surrounding the moving pla

Pregunta (español): ¿Por qué se utiliza la CNN?
Respuesta: universidad de valladolid escuela tecnica superior de ingenieros de telecomunicacion telematica  simultaneous localization an

Pregunta (español): ¿Qué funciones de pérdida se probaron?
Respuesta:  [CLS]   integrated systems in vehicles are essential for sensing the environment surrounding the moving platform, it means t

Pregunta (español): ¿Cómo se realiza la inferencia sobre el conjunto de test?
```

```
Respuesta:  resumen la investigacion sobre el desarrollo de vehiculos autonomos esta evolucionando rapidamente y, con ello, las tecnolog

Pregunta (español): ¿Cuál es la métrica empleada para medir el rendimiento de los modelos?
Respuesta:    simultaneous localization and mapping ) desarrolladas para hallar la localizacion y creacion de mapa del entorno al mismo

Pregunta (español): ¿Cómo se puede interpretar los modelos?
Respuesta: [CLS]  el vehiculo pueda percibir el entorno y actue en consecuencia  integrated systems in vehicles are essential for sensin

Pregunta (español): ¿Cuáles podrían ser las conclusiones del trabajo?
Respuesta:  [CLS] las posibilidades que ofrece uno de estos sistemas anadiendo recursos necesarios para cumplir las finalidades propuest
```

- ¿Cuántos lenguajes puede manejar el BERT para resolver preguntas? BERT (Bidirectional Encoder Representations from Transformers) en sí mismo no está limitado a un conjunto específico de lenguajes, ya que su enfoque de entrenamiento es predecir palabras en un contexto bidireccional.