Sergio Hinojosa
LLM – Especialización en IA FIUBA

LLM con T5

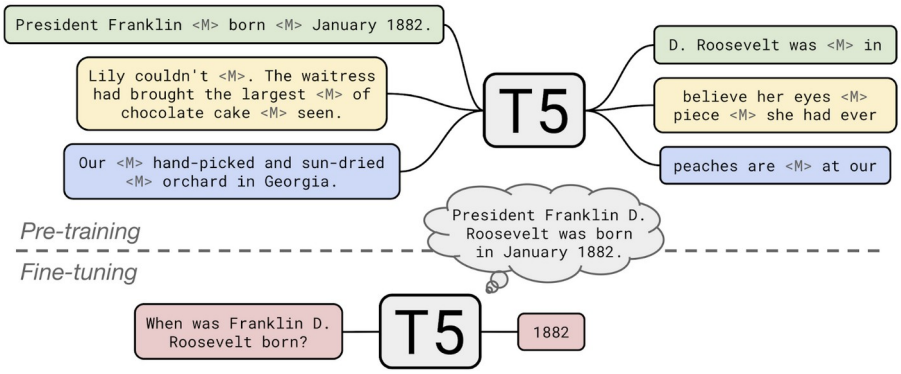
Entrenamiento de un modelo T5 para la descripción de patologías médicas.

.UBAfiuba
FACULTAD DE INGENIERÍA

1 Introducción

1.1 Acerca de T5

T5 es un modelo encoder-decoder que convierte todas las tareas de procesamiento del lenguaje natural (NLP), como la creación de resumenes, traducción, interacciones pregunta-respuesta, generación de texto, etc., en una tarea de secuencia a secuencia, es decir, convierte una secuencia de texto (texto de origen) en otra secuencia de texto (texto objetivo).



Los modelos T5 están previamente entrenados en Colossal Clean Crawled Corpus (C4)[1], que contiene texto y código extraídos de Internet. Este proceso de preentrenamiento permite a los modelos aprender habilidades generales de generación y comprensión del lenguaje. Luego, los modelos T5 se pueden ajustar posteriormente para tareas más específicas.

1.2 El problema

A partir de preguntas sobre patologías y sintomás se propone entrenar un modelo capaz de describir la patología requerida.

1.3 El dataset

El dataset proviene de MedQuAD: un conjunto de datos de respuesta a preguntas médicas.

MedQuAD incluye 47,457 pares de preguntas y respuestas médicas creados a partir de 12 sitios web de los NIH (por ejemplo, cancer.gov, niddk.nih.gov, GARD, MedlinePlus Health Topics). La colección cubre 37 tipos de preguntas (por ejemplo, Tratamiento, Diagnóstico, Efectos secundarios) asociadas con enfermedades, medicamentos y otras entidades médicas como pruebas.

question	**answer**
What causes Heart Failure ?	Heart failure is caused by other diseases or conditions that damage the heart muscle such as coronary artery disease (including heart attacks) …
Who is at risk for Heart Failure? ?	Preventing Heart Failure There are a number of things you can do to reduce the risk for coronary artery disease and heart failure ...

https://www.kaggle.com/datasets/jpmiller/layoutlm/data

1 C4: https://github.com/google-research/text-to-text-transfer-transformer?tab=readme-ov-file#c4

2 Desarrollo

El core del sistema es el modelo y el tokenizador preentrenado.

En este trabajo se utlizó la librería Transformers. La librería Transformers de Python, desarrollada por Hugging Face, es una biblioteca opensource que proporciona una interfaz para trabajar con modelos de lenguaje basados en Transformers, como BERT, GPT, o T5, como en este trabajo.

```
from transformers import (T5ForConditionalGeneration, T5TokenizerFast as T5Tokenizer)

tokenizer = T5Tokenizer.from_pretrained("t5")
model = T5ForConditionalGeneration.from_pretrained("t5", return_dict=True)
```

Se utilizó pytorch_lightning como framwork para la configuración y control del modelo.

2.1 Entrenamiento

Se configura dos objetos principales: El modelo y el data module.

El data modulo se configura, el tamaño de los batchs de entrenamiento, la cantidad de tokens que ingresan al transformer y los tokens que se obtendran del decoder:

```
batch_size (int): batch size.
source_max_token_len (int): max token length of source text.
target_max_token_len (int): max token length of target text.
```

En el entrenamiento del ejemplo la configuración del entrenamiento es el siguiente:

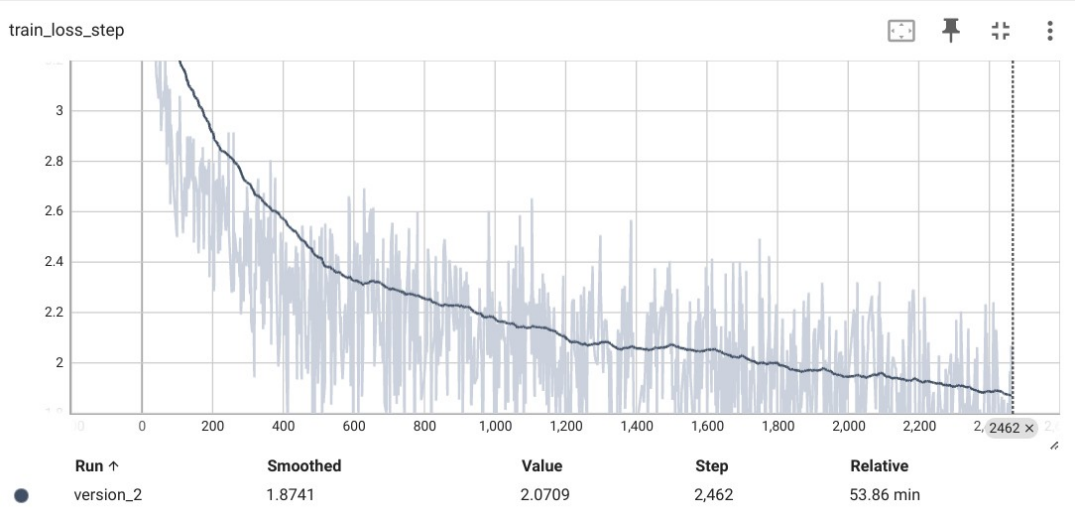
```
import t5

model = t5.SimpleT5()
model.from_pretrained(model_type="t5", model_name="t5-base")
model.train(train_df=train_df,
            eval_df=test_df,
            source_max_token_len=50,
            target_max_token_len=150,
            batch_size=16,
            max_epochs=3,
            use_gpu=True,
            save_only_last_epoch=True)
```

La salida del entrenamiento genera 2 carpetas outputs y lightning_logs.

En outputs se guarda el modelo ya acondicionado con el nuevo dataset y listo para poder inferir información.

En lightning_logs se recolecta la información del progreso del entrenamiento. Se visualizan estás métricas con tensorboard:

```
tensorboard --logdir "./lightning_logs"
```



Se puede ver como en pocas epochs se logra converger la perdida, el modelo es bastante eficiente en el aprendizaje.

2.2 Inferencia

Algunos ejemplos de inferencia. Las entradas son en inglés ya que la semantica del preentrenamiento del modelo está en este lenguaje.

```
[12] input_text="What can cause heart failure?"
     model.predict(input_text)
```

```
["Heart failure is caused by a combination of factors, including genetic and environmental factors. The most common cause of heart failure is diabetes mellitus (HDM). Diabetes mellitus is a condition that affects the heart's ability to pump blood through the body. In some cases, diabetes mellitus can be caused by mutations in the SLC1A gene. Other causes may include high blood pressure, low blood sugar, and/or hypertension."]
```

```
[13] input_text="Can you tell me how to treat Alzheimer's disease?"
     model.predict(input_text)
```

```
["How might Alzheimer's disease be treated? There is no cure for Alzheimer's disease, but it can be treated with medications and lifestyle changes. For example, if you have the condition, your doctor may prescribe medication to treat it. If you have the condition, your doctor may recommend that you take certain medications such as acetaminophen or corticosteroids. Your doctor may also recommend that you take anti-depressants (such as acetaminophen) in combination with other medicines."]
```

```
[14] input_text="symptoms of Breast Cancer"
     model.predict(input_text)
```

```
['Signs and symptoms of breast cancer are caused by breast cancer. Signs and symptoms of breast cancer may be caused by breast cancer or by other conditions. Signs and symptoms of breast cancer may be caused by breast cancer or by other conditions. Signs and symptoms of breast cancer may be caused by breast cancer or by other conditions. Signs and symptoms of breast cancer may be caused by breast cancer or by other conditions.']
```

```
input_text="Are Alopecia and breast pain symptoms of Breast Cancer?"
model.predict(input_text)
```

```
['Alopecia and breast pain are symptoms of breast cancer. They may be caused by a combination of factors, including the presence or absence of certain hormones in the body, such as estrogen, vitamin B12, and other hormones. These and other factors can affect the risk of developing alopecia and breast pain. In some cases, alopecia and breast pain is not a sign of breast cancer. However, it is possible to develop alopecia and breast pain without having any signs or symptoms.']
```

3 Conclusiones

- Buena convergencia: Se pudo observar que el modelo converge muy bien en el aprendizaje en solo 3 epochs.

- Flexibilidad: Se nota la flexibilidad entre la entrada y salida, por ser texto a texto, por lo que no se requiere demasiada especificidad en la estructura del input y target.

- Gran consumo de memoria: Se notó el gran consumo de memoria del entrenamiento por lo cual al ensayarse en Google Colab, se tuvo que probar el modelo con batchs chicos.

- Caja Negra: En estos tipos de modelos se caracteriza por la falta de interpretabilidad, es decir, es díficil interpretar y entender la toma decisiones.

- Tiempo de entrenamiento: Estos modelos son fuertes gracias a la gran cantidad de datos incluidos en el entrenamiento, en este ensayo la esecificidad se dío con un dataset de solo 13000 entrada, con solo 3 epochs el entrenamiento duró alrededor de 1:00hs.

- Transferencia de conocimiento: Al ser entrenado en una amplia gama de tareas de NLP, el conocimiento adquirido por el modelo en una tarea puede transferirse a otras tareas relacionadas. Esto hace que el modelo sea eficiente en términos de recursos y tiempo de entrenamiento. En este caso agregarle especificidad hacia información médica fue exitoso con un dataset no muy grande.