# The Transformer

**Inteligencia Artificial - Ingeniería del Software**
**Propuesta de trabajo**

Alberto Rincón Borreguero

Curso 2021/2022

## 1 Introducción

El Transformer [1] es una arquitectura de red neuronal cuya potencia reside en los mecanismos de atención. Este tipo de redes, propuestas en 2017, han supuesto una revolución en el tratamiento de datos secuenciales. Sus resultados han establecido el estado del arte en varias tareas de procesamiento del lenguaje natural e.g.: traducción automática (*machine translation*), en gran parte debido al artículo publicado por investigadores de Google AI que hacían uso de los transformers en su modelo BERT que servía como base pre-entrenada fácilmente afinable (*fine-tune*) para otros problemas. Posteriormente, desde OpenAI publicaron GPT (actualmente, hay una versión mejorada GPT-3 [2]), cuya repercusión llegó incluso a los medios de comunicación tradicionales debido a su capacidad para generar texto. En este enlace es posible probar una implementación de GPT.

Asimismo, recientemente se ha demostrado la excelente capacidad de los transformers sobre conjuntos de datos compuestos por imágenes. Estas redes neuronales, denominadas Visual Transformers [3], se valen de algunos trucos para adaptar la estructura en 3 dimensiones de las imágenes al formato bi-dimensional y secuencial con el que trabajan los transformers.

En el mundo actual, inmerso en la llamada cuarta revolución industrial, las empresas que quieren ser competitivas tienen en su equipo a personas capaces de entender estos avances en el campo de la inteligencia artificial y de implementarlos en la propia compañía antes de que su desarrollo se estandarice, y por tanto sea fácilmente accesible para todos, dejando de suponer una ventaja competitiva.

## 2 Objetivos

Se espera que con esta propuesta el grupo de trabajo comprenda la arquitectura Transformer y la utilice en una implementación empleando el *framework* PyTorch (sin ayuda de librerías externas para la definición de la red neuronal) que resuelva un problema

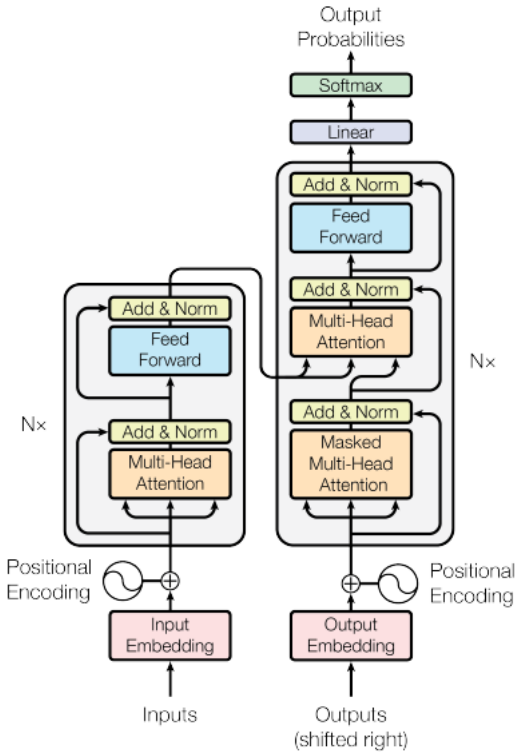


Figure 1: Transformers architecture

propuesto para el dataset proporcionado. Para ello, el grupo de trabajo deberá comprender el funcionamiento básico de PyTorch en su **versión 1.2**. Además, el grupo de trabajo deberá construir otra red neuronal alternativa, de libre elección (recurrente, convolucional, densa, ...) que trate de resolver el mismo problema de forma óptima.

Con este fin, se proponen los siguientes objetivos específicos:

1. Comprender el funcionamiento de la arquitectura Transformer y sus distintas aplicaciones.
2. Ser capaz de extraer las ideas más importantes de artículos científicos.
3. Aprender a usar el framework PyTorch.
4. Llevar a cabo cada uno de los pasos requeridos para el entrenamiento de una red neuronal: obtención, limpieza y preparación de datos; entrenamiento de la red y visualización de errores/métricas; optimización de hiperparámetros.
5. Documentar el trabajo realizado utilizando un formato de artículo científico.
6. Realizar una presentación (PDF o diapositivas) de los resultados obtenidos.

# 3 Descripción

## 3.1 Lecturas e implementación

Para superar los objetivos 1 y 2, se recomienda al grupo de trabajo introducirse en los conceptos básicos mediante la lectura de los artículos de blog: Modelos de secuencia y Transformer: La tecnología que domina el mundo, y, posteriormente, la lectura del artículo científico original *Attention Is All You Need* [1].

Para comprender el funcionamiento de PyTorch y por tanto superar el tercer objetivo, se recomienda al grupo de trabajo completar los 7 pasos de la guía introductoria de la web oficial de PyTorch. Se podrá profundizar mucho más a través del libro Deep Learning with Pytorch, disponible en abierto.

A cada grupo de trabajo le será proporcionado un conjunto de datos sobre el que tendrá que realizar la tarea propuesta, pudiendo ser ésta de clasificación o generación de imágenes, texto o audio. Será decisión del grupo de trabajo la división del conjunto de datos.

Para acelerar el entrenamiento de las redes neuronales construidas, se recomienda el uso de Google Colaboratory que proporciona acceso gratuito a GPUs, y TPUs. Existen otras alternativas como Kaggle (también recomendada), Paperspace, etc.

## 3.2 Documentación y entrega

La memoria de trabajo deberá seguir el formato de artículo científico con una extensión mínima de 6 páginas. Se valorará el uso de LaTeX.

El contenido de la memoria debe ser el siguiente:

- Introducción breve que ponga en contexto al lector.
- Descripción del Transformer así cómo de cualquier otra técnica empleada.
- Descripción de los modelos construidos. Se deberán indicar las dificultades encontradas durante la implementación así como las decisiones tomadas al respecto.
- Presentación de los resultados, incluyendo una comparativa entre las dos redes implementadas.
- Conclusiones.
- Bibliografía.

La entrega del trabajo consistirá en la memoria en formato PDF, el código implementado y un fichero .pth por cada modelo entrenado. El modelo elegido por el grupo de trabajo como aquel que mejor resuelve el problema propuesto deberá llamarse *modelo_seleccionado.pth* para que esta parte sea evaluada. Tanto la memoria como el código deberán subirse a la web de la asignatura en un **único** fichero **.zip**. Se valorará muy positivamente el uso de Github como control de versiones del código desarrollado. Los ficheros .pth se subirán a Github, Google Drive o similar, proporcionando los permisos y el enlace correspondiente.

### 3.3 Presentación y defensa

Como parte de la evaluación del trabajo se deberá realizar una defensa del mismo, para lo que se citará a los alumnos de manera conveniente. El día de la defensa se deberá realizar una pequeña presentación (PDF, PowerPoint o similar) de 10 minutos en la que participarán activamente todos los miembros del grupo que ha desarrollado el trabajo. Esta presentación deberá seguir a grandes rasgos la misma estructura que la memoria del trabajo, haciendo especial mención a los resultados obtenidos y al análisis crítico de los mismos. En los siguientes 10 minutos de la defensa, el profesor procederá a realizar preguntas sobre el trabajo, que podrán ser tanto de la memoria como del código fuente.

## 4 Evaluación

La evaluación del trabajo únicamente procederá si se han desarrollado al menos 2 redes neuronales que resuelvan el problema. Una de ellas debe utilizar Transformers mientras que la otra es de libre elección, considerándose positivamente el uso de redes recurrentes o convolucionales.

Se tendrán en cuenta los siguientes criterios:

- Memoria del trabajo (hasta 1 punto):se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje. La elaboración de la memoria debe ser original, por lo que no se evaluará el trabajo si se detecta cualquier copia del contenido.

- Código fuente (hasta 1 punto): se valorará la claridad y buen estilo de programación, uso de git junto con Github para control de versiones, corrección y eficiencia de la implementación y calidad de los comentarios. El código debe ser original, por lo que no se evaluará el trabajo si se detecta código copiado o descargado de internet.

- Modelo seleccionado (hasta 1 punto): se valorará tanto de manera absoluta como comparativamente con el resto de trabajos que hayan utilizado el mismo conjunto de datos. Para ello, el grupo de trabajo debe exportar el modelo a modelo.pth (y modelo_seleccionado.pth) e incluirlo en la entrega final.

- Presentación y defensa (hasta 1 punto): se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo así como, especialmente, las respuestas a las preguntas realizadas por el profesor.

**IMPORTANTE**: cualquier plagio, compartición de código o uso de material que no sea original y del que no se cite convenientemente la fuente, significará automáticamente la calificación de cero en la asignatura para todos los alumnos involucrados. Por tanto, a estos alumnos no se les conserva, ni para la actual ni para futuras convocatorias, ninguna nota que hubiesen obtenido hasta el momento. Todo ello sin perjuicio de las correspondientes medidas disciplinarias que se pudieran tomar.

## References


[1] Ashish Vaswani et al. "Attention Is All You Need". In: *NIPS2017* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[2] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165.

[3] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *CoRR* abs/2010.11929 (2020).