Una visión general didáctica de las aplicaciones de Transformer: variantes de modelos y directrices

M. I. Cabrera-Bermejo[1], M. Germán[1], D. de la Rosa[1], M. J. del Jesus[1], A. J. Rivera[1], D. Elizondo[2], F. Charte[1], and M. D. Pérez Godoy[1]

[1]Intelligent Systems and Data Mining Group, Department of Computer Science, Univeristy of Jaén, Jaén, Spain. {mbermejo, mgerman, drrosa, mjjesus, arivera, fcharte, lperez}@ujaen.es

[2]Department of computer Technology, De Montfort University, Leicester. UK. elizondo@dmu.ac.uk

Resumen. El presente keywork constituye un resumen del trabajo titulado "A didactic overview of Transformer applications: model variations and user guidelines", publicado en la revista *Logic Journal of the IGPL* en 2026 [1]. Este estudio propone una taxonomía estructurada y una perspectiva didáctica sobre las aplicaciones de los Transformers según el tipo de dato de entrada. Además, se incluyen recomendaciones prácticas que ayudan a guiar el proceso de selección, desarrollo y aplicación correcta de estos modelos de acuerdo a la naturaleza de los datos.

Keywords: Redes Neuronales · Transformers · Aplicaciones.

1 Introducción

La relevancia actual de los modelos basados en la arquitectura Transformer ha impulsado su uso en multitud de campos y en el procesamiento de diversos tipos de datos para resolver tareas muy dispares. En la actualidad, un número creciente de sistemas de Inteligencia Artificial (IA) emplean modificaciones de dicha arquitectura como núcleo fundamental, destacando ejemplos ampliamente reconocidos tanto en la comunidad científica como en la sociedad, tales como la familia de modelos GPT [3], BERT [4], Bart [5], LLaMA [6] o Gemini [7].

Originalmente, el modelo Transformer [2] fue diseñado para la tarea de traducción automática. Su aplicación se ha extendido significativamente para manejar diversas modalidades de información, siendo necesaria la modificación de su arquitectura para optimizar el rendimiento de acuerdo al tipo de dato al que se aplica. La proliferación de nuevas arquitecturas plantea un desafío importante para detectar y comprender sus diferencias respecto al modelo original.

Las revisiones sistemáticas en la literatura actual se centran en el procesamiento de un tipo específico de dato o en un área de aplicación concreta. Nuestro trabajo proporciona una visión global de los Transformers y ofrece directrices generales de uso según el tipo de problema. El objetivo de este *keywork* es difundir las principales aportaciones del trabajo: una taxonomía de aplicaciones y una serie de directrices y recomendaciones clave en el uso de Transformers.

2 Taxonomía de aplicaciones de Transformers

Se propone una taxonomía (Fig. 1) fundamentada en la naturaleza de los datos de entrada y aplicaciones de los modelos basados en arquitecturas Transformer [1]. En dicha taxonomía, se identifican cinco grandes categorías de datos asociadas a tareas de aplicación específicas.

Datos textuales. Representan el dominio nativo de los Transformers, abarcando tareas tradicionales como la traducción automática, la clasificación, la generación de secuencias de texto y los enfoques multitarea. Las variantes se orientan al desarrollo de arquitecturas y mecanismos de atención que permitan reducir la complejidad computacional.

Imagen y vídeo. La adaptación implica la captura de relaciones espaciales y temporales complejas. Sus principales aplicaciones son: clasificación, generación, subtitulado, detección de objetos y sistemas multitarea. A nivel modular, se observa una tendencia a sustituir la función de activación ReLU por GELU en las redes *feed forward* (FFN) y a alterar la disposición de las capas de normalización.

Audio. Este enfoque se centra en las tareas de reconocimiento de voz, separación de fuentes, generación, mejora de la señal y traducción. Las modificaciones más habituales consisten en la eliminación de la codificación posicional absoluta o su reemplazo por codificaciones relativas o capas convolucionales previas.

Datos tabulares o estructurados. Comprende el procesamiento de conjuntos de datos heterogéneos, ya sean secuenciales o no, para resolver problemas de clasificación, regresión y detección de anomalías. Las variantes estudian cómo alterar el cálculo de la atención para capturar correlaciones entre variables, así como *positional encodings* alternativos.

Datos multimodales. Diseñados para procesar e integrar simultáneamente múltiples fuentes de información de diversa naturaleza con el fin de resolver problemas de clasificación, generación y aproximaciones multitarea. En la literatura se observa una clara distinción entre los enfoques que incorporan un único Transformer integrado con un proceso previo de fusión de los *embeddings*, y aquellos que implementan el acoplamiento de múltiples Transformers independientes mediante mecanismos de atención cruzada (*cross-attention*).

3 Directrices y recomendaciones de uso

La adaptación de la arquitectura Transformer a un nuevo dominio o problema específico requiere tomar decisiones de diseño precisas sobre sus componentes internos. Como se muestra en el esquema de la Fig. 2, el proceso requiere la evaluación de cuatro dimensiones estructurales fundamentales para garantizar un rendimiento óptimo y un uso eficiente de los recursos computacionales:

Arquitectura general. La elección de la topología base está ligada a la naturaleza de la tarea. Los modelos compuestos únicamente por el codificador son la opción preferida para tareas de clasificación, comprensión o extracción de características. Por otro lado, las arquitecturas basadas exclusivamente en el decodificador dominan el ámbito de la generación de contenido secuencial.



Fig. 1. Taxonomía de las principales aplicaciones de los Transformers según el tipo de datos de entrada utilizado (adaptado de [1]).

En última instancia, la estructura completa se reserva predominantemente para abordar problemas de traducción o transformación compleja de secuencias de entrada a secuencias de salida.

Mecanismo de atención. El cálculo de atención original propuesto en el artículo [2] exhibe una eficacia notable. Sin embargo, su complejidad computacional cuadrática lo hace ineficiente al procesar secuencias extensas. En contextos como imágenes de alta resolución, fragmentos largos de audio o documentos de texto extensos, resulta crucial sustituirlo por mecanismos de atención lineal, aproximaciones dispersas (*Sparse*) o cálculos de atención personalizados adaptados a la topología específica de los datos (como la atención por ejes espaciales).

Positional encoding. Se recomienda el uso de codificación absoluta para secuencias finitas y cerradas. La codificación relativa es apropiada cuando las distancias entre los elementos son más relevantes que su ubicación precisa, siendo una mejora habitual en el procesamiento de audio o lenguaje natural. Asimismo, existen casos donde es preferible diseñar una a medida para modelar grafos o estructuras tridimensionales, o directamente prescindir de ella si los datos carecen de orden inherente o si se emplean capas convolucionales previas.

Redes *feed forward* (FFN). Las opciones de diseño varían desde la implementación de topologías personalizadas hasta su completa omisión en arquitecturas que persiguen una ligereza extrema. Dentro de este bloque, una de las decisiones más críticas es la selección de la función de activación. De acuerdo con la literatura reciente, se ha producido una clara transición desde la tradicional ReLU hacia variantes más suaves y estables como GELU. Esta modificación ha demostrado ser especialmente determinante para estabilizar el entrenamiento en modelos de visión artificial y sistemas multimodales.

4 Conclusiones

La adopción y expansión de la arquitectura Transformer en dominios ajenos al procesamiento del lenguaje natural ha dado lugar a la generación de un ecosistema complejo de variantes y modificaciones. En el presente *keywork*, se han

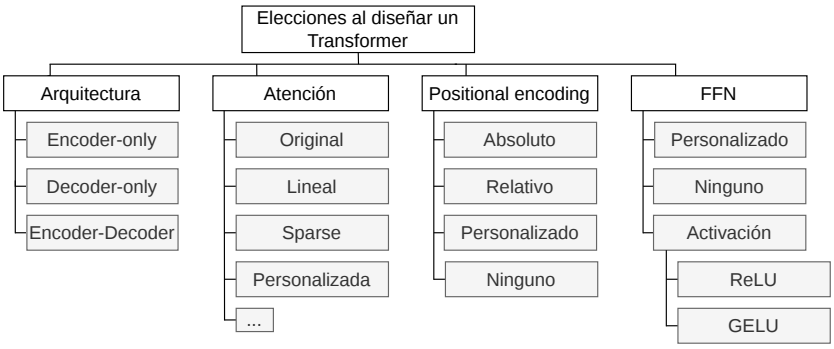


Fig. 2. Principales directrices de diseño y configuración de los componentes internos de un Transformer según las necesidades del problema (adaptado de [1]).

sintetizado las contribuciones fundamentales del estudio original [1], proporcionando a la comunidad investigadora una hoja de ruta estructurada para abordar la selección y el diseño de dichos modelos. Mediante la taxonomía presentada, basada en la naturaleza de los datos de entrada, y de las directrices de diseño estructuradas en cuatro dimensiones críticas (arquitectura, atención, codificación posicional y redes FFN), este trabajo ofrece una guía didáctica y práctica, facilitando la toma de decisiones orientadas a optimizar el rendimiento y la eficiencia computacional de los Transformers en futuros desarrollos.

Agradecimientos. Esta investigación forma parte del proyecto "Advances in the development of trustworthy AI models to contribute to the adoption and use of responsible AI in healthcare (TAIH)" con código PID2023-149511OB-I00, y ha contado con el apoyo del programa de ayudas FPU (Formación de Profesorado Universitario), ambos financiados por el Ministerio de Ciencia, Innovación y Universidades de España.

Referencias

1. Cabrera-Bermejo, M.I., Germán, M., de la Rosa, D., et al.: A didactic overview of Transformer applications: model variations and user guidelines. Logic Journal of the IGPL **34**(1), jzaf040 (2026).
2. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017).
3. OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023).
4. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT, pp. 4171–4186 (2019).
5. Lewis, M., Liu, Y., Goyal, N., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the ACL, pp. 7871–7880 (2020).
6. Touvron, H., Lavril, T., Izacard, G.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
7. Google DeepMind: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023).