Jerónimo Molina Molina

Procesamiento del Lenguaje Natural

5. Introducción a Transformers en PLN

Introducción a Transformers en PLN

Ya conoces las arquitecturas Transformer. Las has estudiado en *Machine Learning II*.

No vamos, pues, a profundizar en su arquitectura, pero sí que repasaremos algo de información básica acerca de esta arquitectura que está revolucionando todo (y quiero decir **TODO**) en el mundo IA.

¿Qué es un modelo transformer?

Es un modelo de aprendizaje profundo que tiene la capacidad de aprender un contexto (significado) siguiendo relaciones en datos secuenciales (como por ejemplo un texto). De ahí que resulten tán útiles en NLP y que los modelos del lenguaje alcancen todo su potencial cuando el sistema de almacenamiento es una base de datos vectorial con soporte de grafos (los nodos de un grafo pueden ser diferentes vectores de chuncks de documentos, pero también desde una perspectiva superior, pueden ser diferentes documentos, o incluso resúmenes de documentos vinculados a otros documentos...).

De hecho, podríamos afirmar que esta arquitectura nos ha permitido empezar a hablar en serio de NLU y olvidarnos un poco de NLP. Aunque obviamente, sin NLP nunca podríamos referirnos a NLU, por lo que debemos entender esta afirmación de forma muy relativa. Pero esta arquitectura, sin duda, ha cambiado las reglas del juego en el procesamiento del lenguaje natural.

Introducción a los Transformer

Sabes que ese elemento diferencial son las capas de atención, un conjunto de técnicas matemáticas, sobre el que se sigue trabajando hoy en día, y que tiene su origen en el artículo publicado en Diciembre de 2017 "***Attention is all you need***" (Google) y que inicia una nueva línea de investigación.

En el siguiente enlace puedes acceder al artículo original: https://arxiv.org/abs/1706.03762, que dada su relevancia, se ha incorporado a la bibliografía esencial de la asignatura y se pone a tu disposición en formato PDF off line.

Pero,…

¿Por qué se ha impuesto esta arquitectura?

Dicho de otro modo, ¿cuál es el valor real de las capas de atención?

Bien, otros modelos que ya hemos estudiado, como las redes recurrentes (RNN) disponen de memoria a corto plazo. Lo mismo les ocurre, aunque en menor medida a las LSTM y redes GRU *-unidades recurrentes cerradas-*, es decir, tienen memoria limitada.

¿Significa esto que debemos olvidarlas? No, estas arquitecturas pueden ser soluciones de bajo coste (computacional) para problemas concretos, y ofrecer en estos casos, soluciones con mayor rendimiento que una arquitectura transformer. Esto puede ser muy útil, por ejemplo, en "IA on Edge", donde instalar IA's en dispositivos pequeños, como FPGA's puede presentar obstáculos relacionados con el rendimiento o los recursos, y por ejemplo, en ocasiones es necesario recurrir al prunado o poda de modelos de aprendizaje profundo.

Sin embargo, la arquitectura transformer, que si, que es verdad que presenta mayor complejidad computacional, ofrece, con sus capas de atención, una ventana de referencia potencialmente infinita, que facilita una comprensión contextual mucho más rica. Este es el valor añadido y el cambio de foco de las arquitecturas transformer frente a enfoques anteriores.

¿Cuál es el siguiente paso?

¿Te has preguntado qué hay más allá de las arquitecturas Transformer?

Bueno, yo soy de los que piensan que los transformer son un gran paso dentro de la IA, y que tienen todavía mucho que ofrecer, si bien, estoy convencido que no nos van a conducir a generar una IA fuerte, que para ello, además de emplear *quantum-ML* (Aprendizaje Automático adaptado a Computación Cuántica) se necesita un enfoque de arquitectura nuevo, mucho más generalista a nivel de contexto y concepto, y que tenga la capacidad de relacionar conocimiento multitarea, además de dotar al sistema de conciencia, consciencia y capacidad de auto-aprendizaje, pero es verdad que para eso aún queda mucho y que con los transformer hay un largo recorrido que nos reportará conocimiento.

Prueba de ello es que se está investigando en dotar del menos común de los sentidos -sentido común- a los modelos basados en Transformers. Esto va a suponer, de hecho, está suponiendo, un gran avance en Natutral Language Understanding. En esta línea, se recomienda la lectura del TFG de un alumno de matemáticas de la universidad de Cantabria, ***Diego Rivera López-Brea***, "Attention is not all you need" (sept-2022) que aborda la ***inferencia causal*** y el ***razonamiento contrafactual*** como líneas de trabajo recientes para dotar a los modelos del lenguaje de sentido común. Como en muchos TFG, el autor realiza un recorrido general -capítulos 1 y 2- del aprendizaje automático, aborda los transformers en el capítulo 3 (arquitectura, buen momento para repasarla), se sumerge en un ejemplo en el capítulo 4 y se centra en el capítulo 5 profundiza en las líneas de investigación mencionadas arriba.

Te recomiendo, para completar esta introducción, la lectura del siguiente artículo, que te dará una visión general de la importancia de los Transformers en el mundo de la inteligencia artificial: https://www.linkedin.com/pulse/transformers-attention-all-you-need-fredy-silva-o-/?originalSubdomain=es.

Aplicación de Transformers a NLU

En primer lugar, conviene mencionar las tareas NLU de bajo nivel que deben considerarse para modelos transformer. En concreto, diferentes organizaciones, entre las que se encuentran Google DeepMind y la Universidad de Washington, han diseñado un test, llamado SuperGLUE (*G*eneral *L*anguage *U*nderstanding *E*valuation) que marca estándares en el rendimiento de modelos NLP y sirve como punto de referencia. El objetivo de este marco no es otro que mostrar que, para que un modelo de comprensión del lenguaje resulte de utilidad debe poder aplicarse en diferentes tareas.

En Ar puedes encontrar el artículo relacionado: https://arxiv.org/pdf/1905.00537.pdf

Puedes obtener más información en https://super.gluebenchmark.com/.

En https://super.gluebenchmark.com/leaderboard puedes acceder al ranking…

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
+	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	10	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
+	11	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	12	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	13	Infosys : DAWN : AI Research	RoBERTa-iCETS					91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5

Click on a submission to see more information

Entre otros detalles relevantes, puede observarse que **la inteligencia humana no experta ocupa la posición 8**, lo que evidencia que diferentes modelos Transformer superan en comprensión, según las diferentes tareas del test, a un humano no experto en una materia específica.

El lector también puede observar la puntuación general y el desglose, o puntuación específica de cada modelo para cada una de las tareas (10 en total) que se consideran. Pasemos a describir brevemente cada una de ellas (para profundizar en SuperGLUE te recomiendo que revises a fondo la página):

- BoolQ: Boolean Questions: Es una tarea para la que el modelo debe responder Si o No
- CB: Commitment Bank: En este caso, se entrena al modelo para que, a partir de una premisa y una hipótesis, este sea capaz de decir en qué medida la hipótesis es correcta o se ajusta/relaciona con la premisa.
- COPA: Choice of Plausible Alternatives es una tarea de razonamiento causal, en la que se entrena al modelo con una premisa y dos alternativas posibles
- MultiRC: Multi-Sentence Reading Comprehension es una tarea de respuesta a preguntas tipo True/False a partir de un texto.
- ReCoRD: Reading Comprehension with Commonsense Reasoning Dataset, es un test de detección de respuesta a partir de diferentes opciones. A partir de un texto, se proporciona una pregunta y diferentes entidades que encajan en la pregunta. El modelo identificará la entidad que encaja en la pregunta indicada. Esta entidad aparece enb el texto varias veces, referida de formas diferentes, pero siempre en formas gramaticalmente válidas, por supuesto.
- RTE: Recognizing Textual Entaliment: El modelo debe predecir en qué medida una oración de premisa implica una oración de hipótesis (problema de inferencia del lenguaje natural).
- WiC: Words in Context pone a prueba la habilidad del modelo para procesar una palabra ambigua. El modelo debe analizar dos oraciones y determinar si la palabra en cuestión tiene el mismo significado en las dos oraciones.
- WSC: Winograde Schema Challenge: Determina si el modelo resuelve correctamente problemas de desamiguación. Es una tarea de comprensión lectora, talque en la oración debe identificar, a partir de un pronombre, el sustantivo (de entre una lista) al que se refiere.
- AX-b: Broadcoverage Diagnostics cada ocurrencia del dataset pone a prueba el sentido común, así, dadas dos oraciones, el modelo concluirá si hay relación de implicación o de

contradicción o de neutralidad entre ellas, siendo contradicción o neutralidad la misma opción.

- AX-g: Winogender Schema Diagnostics, diseñado para medir el sesgo de género, cada ejemplo consta de una oración de premisa, un pronombre masculino o femenino y una hipótesis -antecedente del pronombre-.

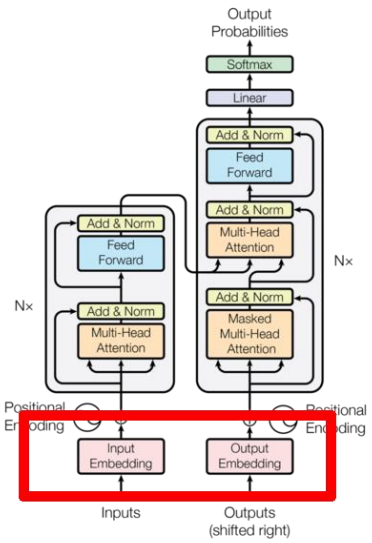
Como puede observarse, SuperGLUE se adentra en la comprensión del lenguaje a través de estos 10 test, en los que el contexto y la gramática desempeñan un papel imprescindible.

A partir de este marco de referencia con tareas de bajo nivel, se presentan diferentes aplicaciones reales de los modelos transformer en NLP/NLU, como por ejemplo pueden citarse:

- Servicios de traducción
- Gestión del conocimiento: Servicios de generación de texto a partir de modelos entrenados con documentación específica. Estos sistemas, son de múltiples aplicaciones:
  - Sistemas de soporte técnico básico
  - Motores de recomendación
  - Repositorios de conocimiento en organizaciones o departamentos
  - etc.
- Detección de emociones
- y muchos otros casos

Ejercicio con Transformers

Teniendo en cuenta la arquitectura de los modelos transformer, ¿te acuerdas de cuál es la entrada a un transformer? Pues si, la entrada es un embedding -aquello que estudiamos en el tema dos y que no es más que un vector multidimensional que representa el contexto del texto al que se refiere-.



De este modo, y sin ánimo de hacer spoiler de temas futuros, imagina el siguiente proceso…

Con esto, quiero transmitirte algo esencial, y es que, en Inteligencia Artificial, todo se basa en conceptos básicos, sencillos, que, si se entienden bien desde el principio, van a servirte, combinados con sentido común, para todo tu futuro en este campo de la ciencia tan interesante como apasionante.

A partir de aquí, vamos a repasar, antes de proponerte un ejercicio de negocio, cada uno de estos pasos:

1. Tokenización: El documento debe trocearse en tokens, que habitualmente son palabras, ya que los modelos del lenguaje y transformers suelen trabajar con ventanas de memoria basadas en palabras, aunque otro token muy utilizado son los caracteres, y no es descartable emplear tokens diferentes en documentos de texto estructurados (títulos, capítulos o secciones de documentos, que, por ejemplo, en formato Markdown son fácilmente identificables)
2. Split: El documento, una vez tokenizado, ya que es, típicamente, de tamaño aceptablemente grande, debe trocearse (chuncking) en porciones lo suficientemente válidas, dependiendo del contexto y del tipo de documento, como para que tengan un significado. Cada uno de estos trozos deben tener un cierto solapamiento, para poderlos relacionar entre sí. Dependiendo del tipo de documento, del sector, etc. el tamaño estándar promedio puede varar; por ejemplo, no es lo mismo trocear imágenes satelitales que trocear un documento de código fuente software o un texto jurídico.
3. Estos Splits, a continuación, se vectorizan, obteniendo un embedding por cada uno de ellos.
4. Finalmente, y hasta aquí puedo leer, a partir de estos embeddings, se realiza un procesamiento, en cada caso, el que corresponda.

Quédate con estos 4 pasos, porque, a partir de aquí, van a ser un proceso que, de forma general, aplicaremos bastante.

En el punto en el que nos encontramos, puedes pensar sin margen de error alguno que esos embeddings son la entrada para cualquier modelo transformer, pero en otros casos, son la entrada para otras tareas.

Propuesta de ejercicio de negocio

Nota: En este ejercicio, puedes elegir si profundizar en la descripción del caso de negocio o bien profundizar en el análisis funcional, dependiendo de si tienes un perfil profesional más de desarrollo negocio o más técnico. Eso sí, profundices en lo que prefieras, hazlo con el máximo nivel de detalle.

Sugerencia: Busca información sobre embeddings y modelos transformer, busca repositorios de modelos transformer ya operativos que puedan dar una solución a las necesidades planteadas en ambos casos.

Descripción del caso 1:

El cliente, una academia de idiomas, te ha encargado que les hagas una propuesta de un sistema basado en Inteligencia Artificial que, empleando transformers, les solucione los siguientes problemas:

- Corrección de ejercicios de traducción -a validar por un profesor finalmente-. La traducción deberá ser inglés-español y español-inglés. El sistema, que recibirá a la entrada un texto

ern inglés, deberá comparar la traducción automática a español con la frase original en español e indicar el nivel de similitud.
- Para textos redactados por los alumnos, necesitan un sistema que realice un análisis sintáctico y muestre el resultado, de tal modo que, de forma rápida, los profesores puedan observar la corrección de las construcciones.

No solo deberás presentar una descripción de negocio o funcional lo más detallada posible, sino que el cliente, que tiene ciertos conocimientos de inteligencia artificial, valorará especialmente que en la propuesta dediques un espacio a explicarle qué modelos transformes vas a emplear y le muestres algún ejemplo o incluso una web donde se vea cómo funcionan.

Descripción del caso 2.

Otro cliente, un centro de documentación regional de Toledo, dispone de un archivo digital inmenso acerca de las diferentes culturas que han habitado en la ciudad, con todo tipo de textos, algunos muy antiguos, traducidos a español moderno.

El cliente solicita un sistema que sea capaz de almacenar toda la información y que permita, mediante interfaz web y de aplicación móvil, responder a preguntas formuladas sobre la documentación digitalizada.

Han contactado contigo para que les sugieras al menos dos posibilidades basadas en IA que puedan ser una solución aceptable a sus propósitos.

Igualmente, desde su área de informática, desean ver modelos de IA operativos que sean de aplicación en las propuestas, que, por otro lado, esperan con todo lujo de detalle funcional y de proceso, porque necesitan estar seguros de que, como profesional, comprender cómo se gestiona todo.

Bibliografía para completar el aprendizaje

1. Libro: Transformers for Natural Language Processing and Computer Vision. Ed. Packt. Aut: Dennis Rothman
2. Artículo: “Attention is all you need”: https://arxiv.org/abs/1706.03762
3. TFG: “Attention is not all you need”: Aut: Diego Rivera López-Brea”, Univ. Cantabria. https://repositorio.unican.es/xmlui/bitstream/handle/10902/26249/RiveraLópez-BreaDiego-TFG-Matemáticas.pdf?sequence=1.
4. Artículo Linked-In: https://www.linkedin.com/pulse/transformers-attention-all-you-need-fredy-silva-o-/?originalSubdomain=es
5. Artículo de blog de nVidia: https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/. (Transformers y aplicaciones avanzadas).
6. https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/
7. Artículo sobre el Test SuperGLUE: https://arxiv.org/pdf/1905.00537.pdf