Optimización de la Respuesta a Consultas en Posgrado mediante Fine-tuning de Llama 7B con Datos Específicos

1st Juan Carlos Peinado Pereira

Posgrado SOE - UAGRM
jcpeinado@soe.uagrm.edu.bo
Santa Cruz de la Sierra, Bolivia
https://orcid.org/0009-0009-9117-2441

Resumen

El artículo se centra en la propuesta de optimizar la respuesta a consultas dentro de la Unidad de Posgrado de Ciencias de la Computación (SOE) de la Universidad Autónoma Gabriel René Moreno (UAGRM). La investigación surge de los desafíos que enfrenta la SOE en la digitalización de sus servicios y la optimización de la comunicación debido a un elevado volumen de consultas administrativas y académicas, la necesidad de agilizar la difusión de investigaciones y la actualización de los programas académicos, así como la carencia de herramientas especializadas. Para abordar estos desafíos, se propone el desarrollo de un modelo de lenguaje basado en Llama 7B, ajustado mediante fine-tuning con datos específicos de la SOE. El objetivo principal de esta iniciativa es evaluar el impacto de esta optimización en la SOE. Para lograrlo, se han establecido objetivos específicos que incluyen la recopilación y curación de datos relevantes, la realización del fine-tuning del modelo Llama 7B, la evaluación de su rendimiento comparándolo con métodos existentes, y la propuesta de un plan de implementación. El artículo proporciona un contexto relevante y el estado actual del arte en el campo de los modelos de lenguaje grandes (LLMs), mencionando la influencia de las arquitecturas transformer y destacando modelos como BERT, GPT, y especialmente LLaMA y LLaMA2. También se explica el proceso de fine-tuning como una adaptación de modelos preentrenados a tareas o dominios específicos, mencionando estrategias como la transferencia de aprendizaje, el aprendizaje multitarea y el ajuste de instrucciones. En la sección de discusión, se subraya la importancia de la programación avanzada en el desarrollo e implementación de LLMs como Llama 7B, así como en el análisis de datos textuales a través de técnicas como el Modelado de Temas y el clustering. Además, se identifican retos importantes como la recopilación y curación de datos, la evaluación de modelos, la implementación e integración, y la complejidad computacional. Se reconocen limitaciones implícitas, como el enfoque específico del estu-

dio y la dependencia de herramientas y frameworks. Finalmente, el artículo explora posibles líneas futuras de investigación que incluyen la profundización en la aplicación de LLMs ajustados en diversos contextos educativos y otros campos, la investigación en metodologías más eficientes para la preparación de datos, la exploración de diferentes arquitecturas de LLMs y estrategias de fine-tuning, el desarrollo de herramientas que faciliten la implementación de LLMs, la investigación en métricas de evaluación más específicas, y la exploración de la combinación de LLMs con otras técnicas de procesamiento del lenguaje natural. La conclusión del artículo enfatiza el potencial de la propuesta para mejorar la eficiencia y la experiencia de la comunidad educativa mediante sistemas de soporte más inteligentes y automatizados.

Palabras clave: Respuesta a consultas, Herramientas que emplean IA, LlaMa2, Fine tuning, Llms.

Abstract

The article focuses on the proposal to optimize the response to queries within the Computer Science Postgraduate Unit (SOE) of the Autonomous University Gabriel René Moreno (UAGRM). The research arises from the challenges that the SOE faces in the digitalization of its services and the optimization of communication due to a high volume of administrative and academic queries, the need to streamline the dissemination of research and the updating of academic programs, as well as the lack of specialized tools. To address these challenges, the development of a language model based on Llama 7B is proposed, fine-tuned with specific data from the SOE. The main objective of this initiative is to evaluate the impact of this optimization on the SOE. To achieve this, specific objectives have been established that include the collection and curation of relevant data, the fine-tuning of the Llama 7B model, the evaluation of its performance compared to existing methods, and the proposal of an implementation plan. The article provides a relevant context and the current state of the art in the field of large language models (LLMs), mentioning the influence of transformer architectures and highlighting models such as BERT, GPT, and especially LLaMA and LLaMA2. The fine-tuning process is also explained as an adaptation of pre-trained models to specific tasks or domains, mentioning strategies such as transfer learning, multi-task learning, and instruction tuning. In the discussion section, the importance of advanced programming in the development and implementation of LLMs such as Llama 7B is underlined, as well as in the analysis of textual data through techniques such as Topic Modeling and clustering. Furthermore, important challenges such as data collection and curation, model evaluation, implementation and integration, and computational complexity are identified. Implicit limitations are acknowledged, such as the specific focus of the study and the dependence on tools and frameworks. Finally, the article explores possible future lines of research that include further exploring the application of fine-tuned LLMs in various educational contexts and other fields, research into more efficient methodologies for data preparation, exploration of different LLM architectures and fine-tuning strategies, development of tools that facilitate the implementation of LLMs, research into more specific assessment metrics, and exploration of combining LLMs with other natural language processing techniques. The conclusion of the article emphasizes the potential of the proposal to improve the efficiency

and experience of the educational community through more intelligent and automated support systems.

Keywords: Answering queries, Tools that use AI, LLaMA2, Fine-tuning, Llms.

1 Introducción

La Unidad de Posgrado de Ciencias de la Computación (SOE) de la Universidad Autónoma Gabriel René Moreno (UAGRM) enfrenta desafíos significativos en la digitalización de sus servicios y la optimización de la comunicación con su comunidad. La gestión de un alto volumen de consultas administrativas y académicas, junto con la necesidad de agilizar la difusión de investigaciones y la actualización de programas académicos, demanda soluciones innovadoras. La atención al cliente se ve comprometida por la falta de herramientas especializadas que faciliten respuestas rápidas y precisas, mientras que los procesos administrativos carecen de automatización y generación de reportes eficientes.

En este contexto, se propone el desarrollo de un modelo de lenguaje basado en Llama 7B, ajustado mediante fine-tuning con datos específicos de la SOE, para optimizar la respuesta a consultas. Esta solución busca proporcionar respuestas eficientes tanto a usuarios internos como externos, mejorando la experiencia general y liberando recursos administrativos.

El objetivo principal de esta investigación es evaluar el impacto de la optimización de la respuesta a consultas en la SOE mediante el fine-tuning de Llama 7B. Para lograrlo, se plantean los siguientes objetivos específicos: 1) recopilar y curar un conjunto de datos relevante que abarque preguntas frecuentes, documentación normativa y currículos académicos; 2) realizar el fine-tuning del modelo Llama 7B con este conjunto de datos; 3) evaluar el rendimiento del modelo en la respuesta a consultas, comparándolo con métodos existentes; y 4) proponer un plan de implementación para integrar el modelo en un entorno accesible para el personal y los usuarios de la SOE.

2 Contexto y Estado del Arte

Este capítulo proporciona una panorámica de como la programacion avanzada esta innersa en los modelo de lenguajes grandes como los llms, en la arquitectura computacional sobre la que corren estos algoritmos como GPUs y TPUs, y en la forma en que se evaluan y comparan estos modelos con otros metodos de respuesta a consultas. Tambien hablaremos como los frameworks como huggingface y transformers han facilitado el acceso a estos modelos y como se han aplicado en diferentes contextos.

2.1 Topic Modeling

El análisis de texto ha experimentado avances significativos en las últimas décadas y entre las técnicas más destacadas se encuentra el Modelado de Temas(Topic Modeling). Esta metodología, anclada en la minería de texto y el procesamiento de lenguaje natural, ha emergido como una herramienta escencial para descubrir patrones subyacentes y estructuras temáticas en grandes conjuntos de documentos. El Modelado de Temas se adentra en la complejidad de los datos textuales, permitiendo la identificación automática y la organización de tema latentes presentes en un corpus (López, 2024).

El uso de modelos basados en la arquitectura encoder-decoder se ha usado para la gene-

ración de algoritmos de Topic Modeling tanto para textos cortos como para textos extensos. En todos los casos, han mejorado comparado con los resultados generados por herramientas más tradicionales y ampliamente usados como Latent Dirichlet Analysis (LDA) (Blei et al., 2003).

Angelov, (Angelov, 2020) presenta un nuevo modelo que utiliza embeddings semánticos para identificar temas sin requerir parámetros predefinidos ni listas personalizadas. A diferencia de los métodos tradicionales, este captura la semántica tanto de palabras como de documentos, y arroja resultados más informativos así como representativos en los experimentos desarrollados por el autor.

2.2 Clustering

Utilizando una representación en vectores de la información a través de embeddings de texto (Sia et al., 2020) presenta evaluaciones comparativas para la combinación de diferentes embeddings de palabras y algoritmos de agrupamiento al tiempo que analiza el rendimiento bajo reducción de dimensionalidad con PCA1. En algunos casos ha tenido resultados tan sólidos como los obtenidos por modelos de temas clásicos, pero con menor tiempo de ejecución y complejidad computacional. Por esta línea, se encuentra también BERTopic (Grootendorst, 2022) que hace una primera representación de los documentos como text embeddings, realiza un agrupamiento en el espacio de los vectores generados y construye los títulos de los temas utilizando un sistema de representación categórica a través de una tabla TF-IDF2.

2.3 Modelos de Lenguaje

La influencia revolucionaria de las arquitecturas de transformers, introducida en 'Attention is All You Need' (Vaswani et al., 2017), ha sido crucial para los grandes modelos de lenguaje (LLM). Modelos como BERT (Devlin et al., 2019) avanzaron en la comprensión bidireccional del contexto, mientras que la serie GPT (Generative Pre-trained Transformer), especialmente GPT3, con un modelo de 175 mil millones de parámetros y su sucesor GPT4 (Achiam et al., 2023) se mantienen a la cabeza de la innovación y desempeño. Otros de los modelos a destacar son los presentados por MetaAI, LLaMA (Touvron et al., 2023) y LLaMA2 (Touvron et al., 2023), una colección de LLM preentrenados y ajustados finamente que varían en escala desde 7 mil millones hasta 70 mil millones de parámetros y que muestran resultados prometedores en comparación con otros modelos (López, 2024).

2.4 Fine-tuning

El ajuste fino es un proceso en el que un modelo preentrenado se adapta para tareas o dominios particulares continuando con el entrenamiento del modelo utilizando solo un conjunto de datos específico del dominio que es diferente del conjunto de datos original utilizado para entrenar el modelo base. Se utilizan diversas estrategias y enfoques de ajuste fino para ajustar los parámetros del modelo a una necesidad específica. Algunos enfoques de ajuste fino se describen brevemente en este artículo.

1. Transferencia de aprendizaje: En este enfoque, un modelo se inicializa primero con pesos guardados de un modelo preentrenado en un conjunto de datos amplio y general, y luego se entrena posteriormen-

te con datos específicos de la tarea limitados. Los pesos se refieren a los parámetros aprendidos de un modelo que ha sido entrenado en un gran conjunto de datos para una tarea específica, que representan el conocimiento que el modelo ha adquirido durante su proceso de entrenamiento, encapsulando características y patrones relevantes para la tarea para la que fue entrenado originalmente.

2. Aprendizaje multitarea: Aquí, los modelos se ajustan en numerosas tareas relacionadas, aprovechando sus similitudes y diferencias, con el fin de maximizar el rendimiento. Por ejemplo, con un modelo CNN entrenado en un conjunto de datos genérico grande (por ejemplo, KINETICS400), se pueden realizar algunas tareas específicas (por ejemplo, estimar la fracción de eyección del ventrículo izquierdo, la edad del paciente y el sexo del paciente a partir de un ecocardiograma) con un conjunto de datos mucho más pequeño aprovechando las características genéricas que el modelo aprendió del gran conjunto de datos.

3. Ajuste de instrucciones: El ajuste de instrucciones implica ajustar un LLM preentrenado para seguir instrucciones de tareas específicas, como traducción, resumen o respuesta a preguntas. Por ejemplo, en la traducción, el modelo se entrena con ejemplos en los que cada entrada incluye una instrucción como "Traduzca la siguiente oración del inglés al francés", seguida de una oración en inglés y su traducción al francés. Después del ajuste fino, el modelo aprende a seguir las instrucciones de traducción y puede generalizar para traducir nuevas oraciones.

3 Discusion

La programación tiene una influencia significativa en el desarrollo de la investigación, especialmente en campos como el procesamiento del lenguaje natural y la ciencia de la computación, que son centrales en el artículo.

- Oportunidades:

  1. Desarrollo y Aplicación de Modelos Avanzados: La programación avanzada es fundamental para el desarrollo y la implementación de modelos de lenguaje grandes (LLMs) como Llama 7B. Estos modelos, ajustados con datos específicos mediante fine-tuning, ofrecen la oportunidad de optimizar la respuesta a consultas y mejorar la eficiencia en la gestión de información, como se propone para la Unidad de Posgrado SOE.
  2. Facilitación del Análisis de Datos Textuales: Técnicas como el Modelado de Temas (Topic Modeling) y el clustering, que se basan en la minería de texto y el procesamiento del lenguaje natural, son posibles gracias a la programación. Estas metodologías permiten descubrir patrones y estructuras temáticas en grandes conjuntos de documentos, facilitando la investigación en diversas áreas. El uso de embeddings de texto y algoritmos de agrupamiento, implementados mediante programación, también ofrece oportunidades para el análisis de información compleja.
  3. Acceso y Utilización de Herramientas y Frameworks: Frameworks como Hugging Face y Transformers,

mencionados en el artículo, facilitan el acceso y la aplicación de modelos de lenguaje preentrenados. Estos recursos, accesibles a través de la programación, democratizan el uso de la inteligencia artificial en la investigación.

4. Personalización y Adaptación de Modelos: El fine-tuning de modelos preentrenados, un proceso inherentemente ligado a la programación, permite adaptar modelos generales como Llama 7B a tareas y dominios específicos. Esto abre la puerta a la creación de herramientas de investigación altamente especializadas y eficientes. Estrategias como la transferencia de aprendizaje, el aprendizaje multitarea y el ajuste de instrucciones, todas implementadas mediante programación, amplían las oportunidades para la investigación aplicada.

- Retos:

1. Recopilación y Curación de Datos: La programación es fundamental para la extracción, limpieza y preparación de datos necesarios para el fine-tuning de modelos de lenguaje. Este proceso puede ser complejo y requiere habilidades técnicas avanzadas para garantizar la calidad y relevancia de los datos utilizados.
2. Evaluación y Comparación de Modelos: La evaluación del rendimiento de los modelos ajustados y su comparación con métodos existentes es un desafío crítico en la investigación. Diseñar métricas adecuadas, implementar procesos de validación y realizar análisis comparativos rigurosos son tareas que requieren conocimientos sólidos de programación y estadística.
3. Implementación e Integración: La integración de los modelos desarrollados en un entorno accesible para los usuarios finales representa otro desafío. Esto requiere habilidades de desarrollo de software y programación para crear interfaces y sistemas que permitan la interacción con los modelos.
4. Complejidad Computacional: El entrenamiento y la ejecución de modelos de lenguaje grandes como Llama 7B requieren una infraestructura computacional significativa, incluyendo GPUs y TPUs. Esto puede ser una limitación para investigadores con recursos computacionales limitados.

- Limitaciones (implícitas en el contexto del artículo):

1. Enfoque Específico: El artículo se centra en la optimización de la respuesta a consultas en un contexto particular (la Unidad de Posgrado SOE) utilizando un modelo de lenguaje. Si bien destaca el potencial de la programación en este ámbito, no profundiza en otras áreas de investigación donde la influencia de la programación podría ser diferente.
2. Dependencia de Herramientas y Frameworks: La implementación de modelos de lenguaje preentrenados y técnicas de procesamiento de texto a menudo depende de herramientas y frameworks específicos. Si bien estos recursos facilitan el desarrollo de la investigación, también pueden

limitar la flexibilidad y la personalización de las soluciones propuestas.

3. La discusión sobre las técnicas de modelado de temas y clustering se presenta como parte del estado del arte, lo que sugiere que son enfoques existentes que se compararán con la propuesta basada en Llama 7B. Esto podría interpretarse como una limitación en la novedad de aplicar la programación a estos problemas, aunque el enfoque específico con fine-tuning de LLMs busca superar estas limitaciones.

4 Implicaciones para la investigacion doctoral

El artículo destaca la influencia significativa de la programación avanzada en el desarrollo de la investigación, especialmente en campos como el procesamiento del lenguaje natural y la ciencia de la computación. En el contexto de la investigación doctoral, un estudiante podría aprovechar la programación de múltiples maneras según lo expuesto en el artículo. En resumen, la programación se inserta en la metodología de investigación doctoral en este contexto como una herramienta fundamental para el desarrollo, la adaptación, la implementación, el análisis y la evaluación de modelos de lenguaje avanzados como Llama 7B, tal como se propone en el artículo para la optimización de la respuesta a consultas. Un estudiante que planee llevar a cabo una investigación similar dependerá en gran medida de sus habilidades de programación para alcanzar sus objetivos.

5 Conclusiones

El principal hallazgo presentado en el artículo es la propuesta de optimizar la respuesta a consultas en la Unidad de Posgrado SOE de la Universidad Autónoma Gabriel René Moreno (UAGRM) mediante el fine-tuning del modelo de lenguaje Llama 7B con datos específicos de la unidad. Esta iniciativa busca abordar los desafíos en la digitalización de servicios y la optimización de la comunicación dentro de la unidad, con el objetivo de mejorar la eficiencia en la atención al cliente y liberar recursos administrativos. La investigación planea recopilar y curar datos relevantes, realizar el ajuste fino del modelo Llama 7B, evaluar su rendimiento en comparación con métodos existentes y proponer un plan de implementación. El artículo también subraya la influencia crucial de la programación avanzada en el desarrollo e implementación de modelos de lenguaje como Llama 7B, así como en el análisis de datos textuales a través de técnicas como el Modelado de Temas y el clustering. Basándose en el artículo, se identifican varias posibles líneas futuras de investigación5 . Estas incluyen el desarrollo y la profundización en la aplicación de modelos de lenguaje grandes (LLMs) ajustados con datos específicos en el contexto de la educación de posgrado y en otros campos5 . Se plantea la investigación y el desarrollo de metodologías más eficientes y robustas para la recopilación, curación y preparación de conjuntos de datos específicos destinados al fine-tuning de LLMs. También se sugiere la exploración de diversas arquitecturas de LLMs y estrategias de fine-tuning para tareas particulares dentro del ámbito educativo, así como el desarrollo de librerías y frameworks que simplifiquen la implementación e integración de LLMs ajustados en entornos educativos. Además, se menciona la investigación en métricas de eva-

luación más específicas y pertinentes para medir el impacto de los LLMs en la optimización de la respuesta a consultas y otros procesos educativos, y la exploración de la combinación de LLMs con otras técnicas de procesamiento del lenguaje natural como el Modelado de Temas y el clustering. La implementación exitosa del fine-tuning de LLMs para la optimización de la respuesta a consultas en la educación podría marcar un avance hacia sistemas de soporte administrativo y académico más inteligentes y automatizados, mejorando la eficiencia general y la experiencia de la comunidad educativa.

6 Conclusion especifica

Síntesis de hallazgos: El principal hallazgo presentado en el articulo, y discutido previamente, es la propuesta de optimizar la respuesta a consultas en la Unidad de Posgrado SOE de la Universidad Autónoma Gabriel René Moreno (UAGRM) mediante el fine-tuning del modelo de lenguaje Llama 7B con datos específicos de la unidad. Esta iniciativa busca abordar los desafíos en la digitalización de servicios y la optimización de la comunicación dentro de la unidad, con el objetivo de mejorar la eficiencia en la atención al cliente y liberar recursos administrativos. La investigación planea recopilar y curar datos relevantes, realizar el ajuste fino del modelo Llama 7B, evaluar su rendimiento en comparación con métodos existentes, y proponer un plan de implementación. El artículo también subraya la influencia crucial de la programación avanzada en el desarrollo e implementación de modelos de lenguaje como Llama 7B, así como en el análisis de datos textuales a través de técnicas como el Modelado de Temas y el clustering.

Posibles líneas futuras: Basándonos en el artículo, se identifican varias posibles líneas futuras:

- Desarrollo y profundización en la aplicación de modelos de lenguaje grandes (LLMs) ajustados (fine-tuning) con datos específicos en el contexto de la educación de posgrado y en otros campos. La implementación de Llama 7B para la respuesta a consultas en la SOE podría servir como base para explorar el potencial de modelos similares en otras tareas administrativas y académicas, tales como la generación de resúmenes de documentos o la personalización de la información para los estudiantes.

- Investigación y desarrollo de metodologías más eficientes y robustas para la recopilación, curación y preparación de conjuntos de datos específicos destinados al fine-tuning de LLMs. Dada la importancia de la calidad de los datos para el rendimiento del modelo, futuras investigaciones podrían enfocarse en optimizar este proceso, posiblemente a través del desarrollo de herramientas de software especializadas o la aplicación de técnicas de aumento de datos.

- Exploración de diversas arquitecturas de LLMs y estrategias de fine-tuning para tareas particulares dentro del ámbito educativo. Aunque el artículo se centra en Llama 7B, futuras investigaciones podrían comparar el rendimiento de otros modelos o experimentar con diferentes enfoques de ajuste fino, como el aprendizaje multitarea o el ajuste de instrucciones, para identificar las metodologías más efectivas.

- Desarrollo de librerías y frameworks que simplifiquen la implementación e integración de LLMs ajustados en entornos edu-

cativos. El artículo menciona la relevancia de frameworks como Hugging Face y Transformers. Podrían surgir nuevas librerías o extensiones de las existentes que faciliten el proceso de ajuste fino, evaluación e implementación de LLMs para usuarios con menos experiencia técnica en programación.

- Investigación en métricas de evaluación más específicas y pertinentes para medir el impacto de los LLMs en la optimización de la respuesta a consultas y otros procesos educativos. Más allá de las métricas generales de rendimiento de los modelos de lenguaje, se podrían desarrollar métricas que evalúen la utilidad, precisión y satisfacción del usuario con las respuestas proporcionadas por los modelos ajustados.

Exploración de la combinación de LLMs con otras técnicas de procesamiento del lenguaje natural, como el Modelado de Temas y el clustering. Aunque el artículo los presenta como métodos existentes, futuras investigaciones podrían investigar cómo estas técnicas pueden complementar o mejorar el rendimiento de los LLMs en tareas como la identificación de temas relevantes en las consultas o la agrupación de preguntas similares para mejorar las bases de conocimiento. En cuanto a cambios de paradigma, la implementación exitosa del fine-tuning de LLMs para la optimización de la respuesta a consultas en la educación podría marcar un avance hacia sistemas de soporte administrativo y académico más inteligentes y automatizados. Esto podría permitir al personal dedicar menos tiempo a tareas repetitivas y enfocarse en actividades de mayor valor estratégico, mejorando la eficiencia general y la experiencia de la comunidad educativa. La creciente accesibilidad a modelos de lenguaje potentes a través de frameworks y posibles futuras librerías también podría transformar la manera en que se lleva a cabo la investigación en diversas disciplinas, facilitando el análisis de grandes volúmenes de información textual y la generación de nuevas perspectivas.

Referencias

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171-4186.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

López, J. A. O. (2024). *Generación Automática de un Borrador de Estado del Arte a partir de una colección de documentos científicos* [Tesis doctoral, Universidad de La Habana].

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast

and good topics too! *arXiv preprint arXiv:2004.14914*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.