Evaluación de la Red Neuronal Recurrente Mamba para Extracción de Respuestas con PLN

Horacio Brizuela

Trabajo Final del Curso Minería de datos para Texto 2024

FAMAF, Universidad Nacional de Córdoba, Argentina

1 de octubre de 2024

1 Motivación

La extracción de respuestas es una actividad de alta relevancia actual en el campo del Procesamiento del Lenguaje Natural (PLN). Hasta el presente, los modelos de redes neuronales como GPT-3 o BERT (Devlin et al., 2019), basados en transformers, han dominado esta tarea, pero su uso está limitado por sus altos requisitos computacionales (Brown et al., 2020). En contraste, las arquitecturas de las Redes Neuronales Recurrentes (RNNs), como por ejemplo Mamba pueden manejar datos secuenciales con menores recursos (Gu, Dao, 2023) .

Este trabajo evaluará el desempeño de la Red Neuronal Recurrente Mamba para la extracción de respuestas en entornos de recursos limitados.

2 Introducción

Las RNNs pueden ofrecer ventajas sobre modelos como GPT-3 y BERT en entornos donde se necesita optimización de recursos. Las RNNs, como Mamba, podrían ser capaces de mantener el contexto a lo largo de secuencias de texto, y ser útiles en tareas secuenciales.

Mamba is es una arquitectura de deep learning con modelado secuencial. Está basada en el modelo S4 (Structured State Space sequence). Se trata de una red neuronal recurrente que se destaca por su eficiencia en tareas que requieren manejo de contexto, como la extracción de respuestas en textos largos. Su implementación en entornos con recursos computacionales limitados la convierte en candidata para una opción viable y eficiente (Graves, 2013).

3 Hipótesis de Trabajo

Mamba RNN puede realizar la extracción de respuestas de manera efectiva y eficiente en entornos locales. El desempeño de Mamba puede ser comparable a otros modelos grandes en términos de precisión y coherencia.

4 Objetivo Principal

Evaluar el desempeño y características de Mamba en la extracción de respuestas, específicamente su rendimiento computacional y precisión.

5 Actividades

5.1 Instalación y Configuración de Mamba RNN

El primer paso es la instalación y correcta configuración de Mamba RNN en un entorno local, asegurando que dicha configuración sea adecuada para su funcionamiento.

5.2 Determinación del Dataset

Se analizará el tipo de dataset necesario para la extracción de respuestas. Se determinarán los requisitos del dataset y se evaluarán diversas opciones, especialmente aquellas que permitan usar benchmarks.

5.3 Análisis de Características

Se ejecutará el código instalado y se estudiará su funcionamiento. El análisis se centrará en la evaluación de la gestión de datos secuenciales y los mecanismos presentes en Mama. Se investigarán ”scores” y métodos de análisis.

5.4 Evaluación Individual del Desempeño

5.4.1 Estudio de Métricas

Se investigarán métricas para la evaluación de RNNs aplicadas a la extracción de respuestas.

5.4.2 Pruebas y Reportes

Se realizarán pruebas en Mamba y se medirá su desempeño según las métricas. Se generarán reportes basados en estándares.

5.5 Evaluación Comparada del Desempeño

5.5.1 Estudio de Benchmarks

Se investigarán benchmarks adecuados para la evaluación comparativa de RNNs aplicadas a la extracción de respuestas.

5.5.2 Pruebas Comparativas y Reportes

Se realizarán pruebas comparativas entre Mamba y otras redes neuronales, y se medirá su desempeño según las métricas y benchmarks. Se generarán reportes según los estándares establecidos.

6 Planificación

Se divide el volumen total de actividad en los siguientes paquetes de trabajo (Work Packages, WP):

- WP 1 (0,5 meses): Setup Actividad 4.1: Definición de requerimientos para instalación y configuración de Mamba RNN. Actividad 4.2: Determinación del dataset.
- WP 2 (0,5 meses): Análisis Actividad 4.3: Análisis de características.
- WP 3 (2 meses): Evaluación individual del desempeño Actividad 4.4: Evaluación individual del desempeño. Reporte de evaluación individual.
- WP 4 (3 meses): Evaluación comparada Actividad 4.5: Evaluación comparada del desempeño. Reporte de evaluación.

- WP 5 (1 mes): elaboración del reporte final del proyecto. Redacción de artículo científico-tecnológico.

7 Tabla de Actividades y Entregables

ID	Work Package	Duración (meses)	Entregables
1	WP1: Setup	0,5	Reporte de requerimientos y reporte de instalación
2	WP2: Análisis	0,5	Reporte de análisis de características
3	WP3: Evaluación Individual	2	Reporte de evaluación individual
4	WP4: Evaluación Comparada	3	Reporte de evaluación comparada
5	Publicación	1	Reporte final y artículo científico-tecnológico

Table 1: Actividades y Entregables del Proyecto

8 Referencias

- Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Link.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT. Link.
- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems. Link.
- Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. ICLR. Link.
- Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. arXiv preprint arXiv:1308.0850. Link.
- Merity, S., Keskar, N. S., Socher, R. (2017). Regularizing and Optimizing LSTM Language Models. arXiv preprint arXiv:1708.02182. Link.
- Gu, A., Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752. Link.