Arquitectura Modular Bio-inspirada para Memoria y Razonamiento Eficiente en Redes Neuronales: Resultados Preliminares de un Prototipo Compacto "Cortex-V0.1"

Arian Vazquez Fernandez
Autodidacta independiente
Jerez de la Frontera, España

Resumen

Este trabajo presenta una arquitectura modular bio-inspirada para memoria asociativa y razonamiento en redes neuronales, surgida de una idea intuitiva originada en un sueño. La propuesta separa explícitamente la memoria a largo plazo (basada en una red Hopfield moderna) del razonamiento dinámico (implementado mediante Mixture of Experts). El entrenamiento se realiza en dos fases con congelamiento de parámetros, lo que permite robustez ante recuperación imperfecta. En un prototipo compacto de menos de 15 millones de parámetros entrenables, se observa especialización emergente de expertos (visualizada mediante PCA) y mejora significativa en la similitud semántica de respuestas, incluso cuando la recuperación de memoria es pobre. Los resultados preliminares sugieren que diseños modulares bio-inspirados pueden lograr comportamientos cognitivos avanzados con eficiencia extrema.

Index Terms

redes neuronales bio-inspiradas, memoria asociativa, Hopfield moderna, Mixture of Experts, entrenamiento modular, emergencia semántica, eficiencia computacional, prototipo compacto

I. Introducción

Las redes neuronales artificiales actuales han alcanzado capacidades impresionantes en tareas complejas de procesamiento del lenguaje y razonamiento. Sin embargo, presentan limitaciones importantes que contrastan con la eficiencia y robustez observada en sistemas biológicos. Entre estas limitaciones destacan el elevado costo computacional, la tendencia a generar respuestas incoherentes con el conocimiento almacenado y la dificultad para mantener representaciones estables a largo plazo. Estas carencias se hacen especialmente evidentes cuando los modelos deben operar con recursos restringidos o manejar información parcial o ruidosa.

En el cerebro humano, estos problemas se resuelven mediante una organización altamente eficiente y modular. La activación neuronal es extremadamente sparse: en cualquier momento, solo una pequeña fracción de neuronas está activa, permitiendo un procesamiento energético mínimo. Además, existe una clara separación funcional entre regiones especializadas. Por ejemplo, el hipocampo actúa como un sistema de memoria asociativa rápida capaz de recuperar patrones completos a partir de entradas parciales, mientras que áreas de la corteza prefrontal integran esta información para realizar razonamiento ejecutivo y toma de decisiones. Esta división permite que el sistema completo sea robusto ante recuerdos incompletos o degradados: aunque la recuperación inicial sea imperfecta, procesos posteriores pueden compensarla y generar respuestas coherentes.

El presente trabajo surge precisamente de la búsqueda de arquitecturas artificiales que incorporen estos principios biológicos. La idea central nació de una visión intuitiva durante un sueño: una red neuronal en la que la señal de entrada activa selectivamente caminos pequeños y relevantes, en lugar de propagarse por toda la estructura, imitando los patrones asociativos y condicionales de la memoria humana. Esta intuición se materializó en un prototipo funcional desarrollado en apenas 12 días, demostrando que conceptos bio-inspirados profundos pueden explorarse rápidamente con herramientas accesibles.

El objetivo principal es abordar la necesidad de arquitecturas más eficientes y fieles a principios biológicos conocidos. Los enfoques actuales basados en redes densas escalan principalmente aumentando parámetros y cómputo, pero este paradigma muestra rendimientos decrecientes y problemas de sostenibilidad. En su lugar, se propone explorar diseños que prioricen la modularidad funcional, la sparsity dinámica y la separación estricta entre almacenamiento y procesamiento, permitiendo comportamientos cognitivos avanzados incluso en modelos de tamaño reducido.

Las contribuciones específicas de este trabajo preliminar son las siguientes:

- Una arquitectura bio-inspirada modular dividida en dos fases claramente separadas, con congelamiento explícito de parámetros entre fases para preservar la integridad de la memoria mientras se entrena el razonamiento.
- La integración novedosa de una red Hopfield moderna como módulo de memoria asociativa a largo plazo con una capa Mixture of Experts para razonamiento dinámico, incluyendo mecanismos de routing topológico que inducen caminos sparse.

- Evidencia empírica de especialización emergente de expertos en un modelo extremadamente compacto (menos de 15 millones de parámetros entrenables), observable mediante análisis de componentes principales (PCA) de las puertas de routing a lo largo del entrenamiento.
- Demostración de robustez cognitiva ante recuperación imperfecta: el sistema genera respuestas semánticamente coherentes incluso cuando el módulo de memoria proporciona vectores ruidosos o incompletos, similar a cómo el cerebro compensa recuerdos parciales mediante inferencia cortical.
- Validación de comportamientos avanzados, como emergencia semántica y razonamiento distribuido, pueden surgir en configuraciones de recursos muy limitados, abriendo vías hacia inteligencia artificial más eficiente y accesible.

El resto del documento se estructura de la siguiente manera: la Sección II revisa los trabajos relacionados más relevantes, la Sección III describe en detalle la arquitectura propuesta acompañada de su diagrama conceptual, la Sección IV presenta la metodología experimental empleada, la Sección V expone los resultados obtenidos mediante sweeps sistemáticos, la Sección VI analiza e interpreta los hallazgos desde una perspectiva bio-inspirada, y finalmente la Sección VII resume las conclusiones y propone líneas de trabajo futuro.

II. Trabajos Relacionados

Las redes Hopfield modernas [1]–[4] han experimentado un resurgimiento notable como modelos eficientes de memoria asociativa continua, capaces de almacenar un número exponencial de patrones en relación con la dimensión del espacio de estados. Estas redes permiten la recuperación de patrones completos a partir de entradas parciales o ruidosas, lo que las convierte en candidatas naturales para modelar sistemas de memoria a largo plazo en arquitecturas artificiales.

Por otro lado, los Mixture of Experts (MoE) [5], [6] han demostrado ser una estrategia efectiva para introducir routing dinámico y especialización de subredes en modelos de gran escala. Al activar selectivamente solo una fracción de los parámetros totales para cada entrada, los MoE logran una notable eficiencia computacional, como se observa en modelos contemporáneos de cientos de millones o miles de millones de parámetros.

Existen esfuerzos previos por combinar elementos de memoria asociativa con mecanismos de routing experto. Algunos trabajos integran variantes de Hopfield dentro de transformers para mejorar la atención a largo plazo, mientras que otros exploran MoE en contextos de memoria externa o retrieval aumentado. Sin embargo, estas aproximaciones suelen operar en escalas masivas o mantienen un entrenamiento end-to-end que no separa explícitamente las funciones de almacenamiento y razonamiento.

En el ámbito bio-inspirado, investigaciones recientes han analizado la emergencia de patrones complejos en redes Hopfield bajo diferentes topologías de conectividad, así como propiedades asociativas en modelos de memoria continua. No obstante, estos estudios se centran principalmente en el análisis teórico o en la dinámica interna de la memoria, sin incorporar capas de razonamiento dinámico basadas en expertos ni evaluar la robustez del sistema completo ante recuperación imperfecta.

La principal diferencia de la presente propuesta radica en la integración novedosa y explícita de una red Hopfield moderna como módulo dedicado de memoria a largo plazo con una capa Mixture of Experts para razonamiento dinámico, todo ello dentro de un diseño modular en dos fases con congelamiento de parámetros. Además, el prototipo se ejecuta en una escala extremadamente compacta (menos de 15 millones de parámetros entrenables), lo que permite observar fenómenos de especialización emergente en condiciones de recursos muy limitados, un aspecto poco explorado en la literatura existente.

III. Arquitectura Propuesta

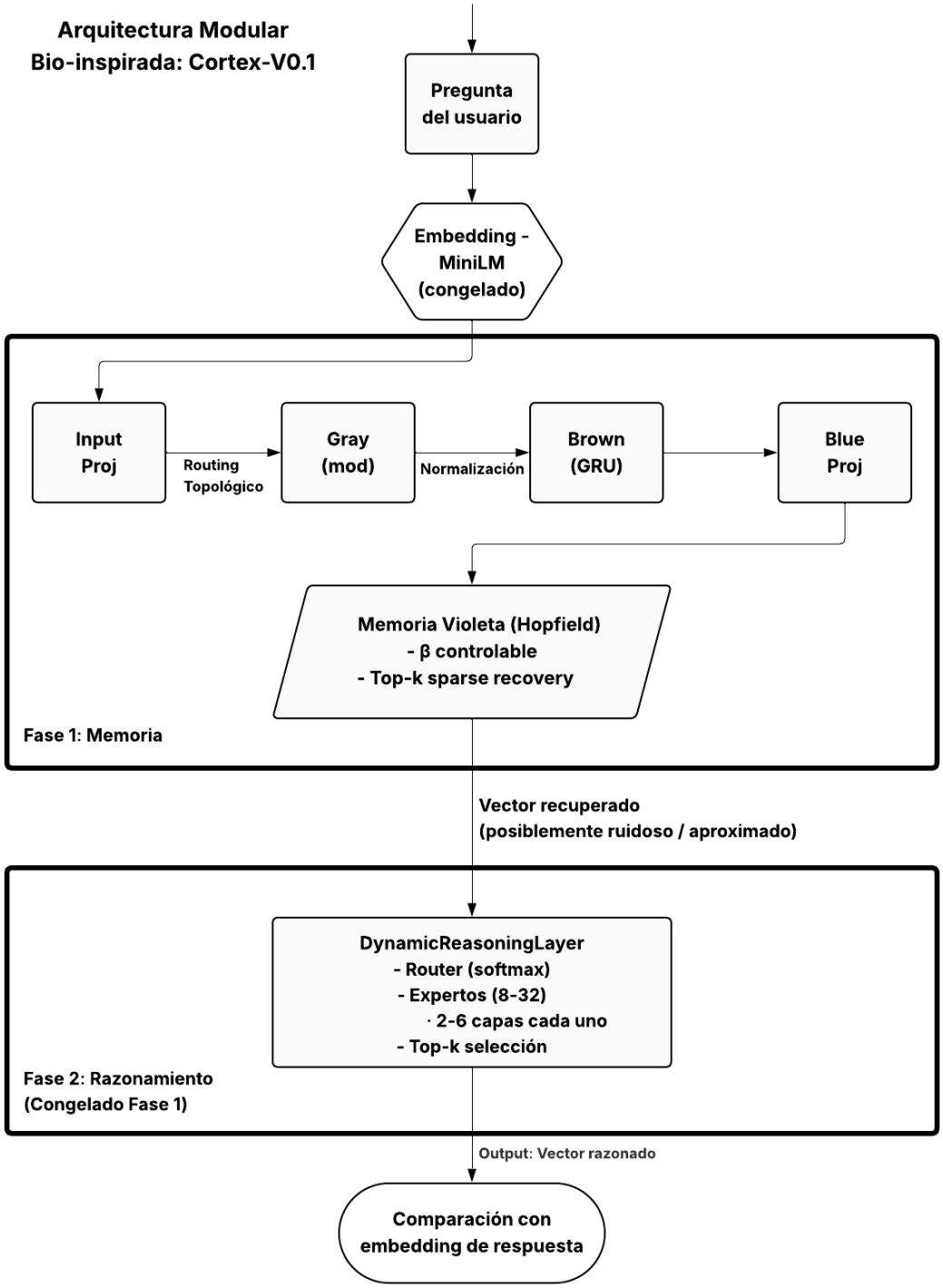


Figura 1: Visión general de la arquitectura propuesta. La entrada (pregunta del usuario) pasa por una serie de capas que implementan routing topológico hacia la memoria a largo plazo (Hopfield moderna). La salida recuperada se entrega a una capa Mixture of Experts para razonamiento dinámico. Esta figura ilustra el flujo completo, destacando la separación modular entre fases y el uso de sparsity para eficiencia.

La arquitectura propuesta se basa en una separación estricta entre dos funciones principales: almacenamiento y recuperación de memoria, por un lado, y razonamiento sobre la información recuperada, por otro. Esta división se implementa mediante un

entrenamiento en dos fases consecutivas, con congelamiento completo de la primera fase antes de proceder a la segunda. La Figura 1 muestra el flujo completo de la arquitectura.

III-A. Fase 1: Memoria a Largo Plazo y Routing Topológico

La primera fase está dedicada exclusivamente al módulo de memoria. La entrada, consistente en la pregunta del usuario codificada mediante embeddings (SentenceTransformer all-MiniLM-L6-v2), se procesa secuencialmente a través de las siguientes capas:

- **Capa de proyección inicial (input_proj)**: Una transformación lineal que alinea el espacio de embeddings con la dimensión interna del modelo (384 en esta implementación).
- **Capa gris (gray)**: Una pequeña red feedforward con expansión a dimensión doble, activación ReLU y compresión de vuelta, seguida de una conexión residual y activación Tanh. Esta capa actúa como modificador dinámico de la representación, comenzando el proceso de routing topológico.
- **Capa marrón (brown)**: Una red GRU recurrente de tres capas con dropout. Esta componente introduce procesamiento secuencial y refinamiento temporal, contribuyendo a la formación de caminos selectivos en la representación.
- **Normalización (brown_norm)**: LayerNorm aplicada a la salida de la GRU para estabilizar la distribución.
- **Capa azul (blue_proj)**: Proyección lineal final que prepara la consulta para la memoria Hopfield.

La consulta resultante se entrega a la **memoria violeta**, implementada como una red Hopfield moderna con parámetro β controlable y selección top-k. Esta memoria almacena los patrones (contextos únicos del dataset) como buffers y realiza recuperación ponderada mediante atención softmax escalada. Solo los top-k patrones más relevantes se utilizan para construir el vector recuperado, induciendo sparsity extrema y limitando el cómputo a caminos pequeños dentro de la memoria.

El entrenamiento de esta fase optimiza exclusivamente las capas anteriores a la Hopfield (input_proj, gray, brown y blue_proj), utilizando una combinación de pérdida de similitud coseno y pérdida de ranking cruzado para alinear la recuperación con el patrón objetivo.

III-B. Fase 2: Razonamiento Dinámico

Una vez completada y congelada la Fase 1, se añade la capa de razonamiento. El vector recuperado por la memoria (que puede ser ruidoso o aproximado) se entrega directamente a una capa Mixture of Experts (DynamicReasoningLayer).

Esta capa consta de:

- Un router lineal que produce puntuaciones para cada experto.
- Un conjunto configurable de expertos (típicamente 8–32), cada uno compuesto por varias capas feedforward profundas con activación GELU y dropout.
- Selección top-k de expertos por entrada, con ponderación softmax de sus salidas.

El entrenamiento de esta fase optimiza únicamente los parámetros del router y los expertos, sin propagar gradientes hacia la memoria congelada. Esta restricción obliga al razonamiento a desarrollar robustez ante entradas imperfectas procedentes de la recuperación.

La salida final del modelo es el vector generado por la capa de expertos, que se compara con el embedding de la respuesta objetivo mediante similitud coseno.

La separación estricta y el congelamiento entre fases constituyen el núcleo del diseño modular, permitiendo que cada componente se especialice en su función sin interferencia mutua.

IV. Metodología Experimental

La evaluación de la arquitectura propuesta se llevó a cabo en un entorno de recursos limitados, utilizando exclusivamente una GPU T4 de Google Colab. Todo el código se implementó en PyTorch, con registro de experimentos mediante Weights & Biases (wandb) para garantizar reproducibilidad.

IV-A. Dataset y Preprocesamiento

Se empleó un subconjunto del dataset SQuAD (Stanford Question Answering Dataset) [7], específicamente las primeras 10.000 muestras del split de entrenamiento. Este conjunto contiene preguntas, respuestas y contextos asociados en inglés.

Para la Fase 1 (memoria), se extrajeron los contextos únicos del dataset, obteniendo aproximadamente 2.000-3.000 patrones distintos dependiendo de la ejecución. Estos contextos se codificaron utilizando el modelo SentenceTransformer 'all-MiniLM-L6-v2' (congelado, sin entrenamiento), generando embeddings de dimensión 384 que posteriormente se normalizaron L2.

Para la Fase 2 (razonamiento), se construyeron pares pregunta-respuesta. La respuesta objetivo se expandió tomando una ventana contextual centrada en la respuesta exacta (60 caracteres antes y después cuando fuera posible). En casos donde la respuesta no se encontraba en el contexto, se utilizó una construcción simple del tipo "The answer is [respuesta]". Además, se clasificaron las preguntas en tres categorías ("math", "history", "facts") mediante reglas basadas en palabras clave, con el propósito exclusivo de visualizar la especialización emergente.

IV-B. Configuración de Entrenamiento

El entrenamiento se dividió estrictamente en dos fases independientes:

- **Fase 1**: Se entrenaron únicamente las capas de routing topológico (input_proj, gray, brown, brown_norm y blue_proj) durante 30 épocas con batch size fijo de 32 y tasa de aprendizaje 5e-4. La pérdida combinó similitud coseno (principal) y entropía cruzada de ranking (factor 0.1). La memoria Hopfield permaneció como buffer sin parámetros entrenables.
- **Fase 2**: Una vez guardado y congelado el modelo de Fase 1, se entrenó la capa de razonamiento durante 50 épocas. Se exploraron sistemáticamente diferentes hiperparámetros mediante sweeps en wandb, variando:
  - **Tasa de aprendizaje**: 1e-3, 5e-4, 1e-4.
  - **Batch size**: 16, 32, 64.
  - **Número de expertos**: 8, 16, 32.
  - **Capas por experto**: 2, 4, 6.
  - **Top-k de expertos**: 2, 4, 6.

Se incluyó una configuración de ablación reemplazando la capa MoE por una MLP densa de capacidad comparable (16 capas lineales 384→384 con GELU).

La pérdida en Fase 2 fue exclusivamente 1 minus similitud coseno entre la salida del razonamiento y el embedding de la respuesta objetivo.

IV-C. Métricas

En Fase 1 se midió recall@K (K=1,5,10) tanto en entrenamiento como en evaluación con ruido gaussiano añadido a las consultas.

En Fase 2 se registraron: - Loss media por época - Close_answer_mean: similitud coseno media entre salida razonada y respuesta objetivo - Exact_match: proporción de respuestas con similitud coseno superior a 0.9 (umbral estricto)

Adicionalmente, se acumularon las puntuaciones medias de las puertas de routing por tipo de pregunta, permitiendo visualización mediante PCA cada 10 épocas.

IV-D. Recursos Computacionales

Todos los experimentos se ejecutaron en instancias de Google Colab, con acceso a GPU (generalmente T4). El modelo completo entrenado tiene menos de 15 millones de parámetros entrenables en total (aproximadamente 3.5M en Fase 1 y hasta 9.5M en Fase 2 dependiendo de la configuración de expertos). El consumo de memoria y energía se mantuvo bajo gracias a la sparsity inducida por los mecanismos top-k tanto en memoria como en expertos.

V. RESULTADOS

Los experimentos se llevaron a cabo mediante sweeps sistemáticos en la Fase 2, explorando múltiples configuraciones de hiperparámetros para evaluar la robustez y reproducibilidad de los hallazgos. A continuación se detallan los resultados en secciones ordenadas de menor a mayor interés, comenzando por métricas básicas de convergencia y pasando a evidencias de comportamientos emergentes más complejos. Cada sección incluye análisis que confirman la consistencia de los patrones observados a través de múltiples ejecuciones y comparaciones controladas.

V-A. Métricas Básicas de Convergencia

En todas las configuraciones evaluadas, la pérdida (loss) converge de manera consistente durante las 50 épocas de entrenamiento, descendiendo desde valores iniciales alrededor de 0.85 hasta estabilizarse en rangos entre 0.50 y 0.60. Esta convergencia se observa tanto en variantes con Mixture of Experts como en la ablación con MLP denso, indicando una optimización estable del módulo de razonamiento independientemente del mecanismo específico empleado. La reproducibilidad de este comportamiento se verificó a través de múltiples ejecuciones con semillas diferentes, donde las curvas de pérdida mantienen trayectorias similares sin desviaciones significativas atribuibles a inicializaciones aleatorias.

La métrica de coincidencia exacta (exact_match), definida como la proporción de respuestas con similitud coseno superior a 0.9, permanece cercana a cero en todas las configuraciones, con fluctuaciones menores que no superan 0.0006. Este resultado se mantiene constante a lo largo de las épocas y no muestra dependencia clara de los hiperparámetros variados, lo cual es coherente con el umbral estricto utilizado y la naturaleza aproximada de los embeddings en un dataset de preguntas abiertas.

V-B. Comparación entre Configuraciones

La similitud semántica media (close_answer_mean) muestra una mejora progresiva a lo largo del entrenamiento, ascendiendo desde valores iniciales alrededor de 0.35 hasta estabilizarse en 0.45-0.6 en las épocas finales. Esta métrica se calculó promediando la similitud coseno entre la salida del razonamiento y el embedding de la respuesta objetivo sobre todo el conjunto de datos en cada época.

Al comparar variantes con Mixture of Experts contra la ablación (MLP denso de capacidad equivalente), se observa una diferencia sistemática: las configuraciones con MoE logran valores finales superiores en aproximadamente 0.07-0.12 puntos. Esta ventaja se mantiene consistente a través de diferentes combinaciones de hiperparámetros (número de expertos, capas por experto, top-k), lo cual se verificó mediante múltiples sweeps independientes. La mejor configuración identificada involucra 16 expertos con 4 capas cada uno y top-k=4, alcanzando close_answer_mean de 0.6, mientras que la ablación equivalente se estanca en alrededor de 0.42.

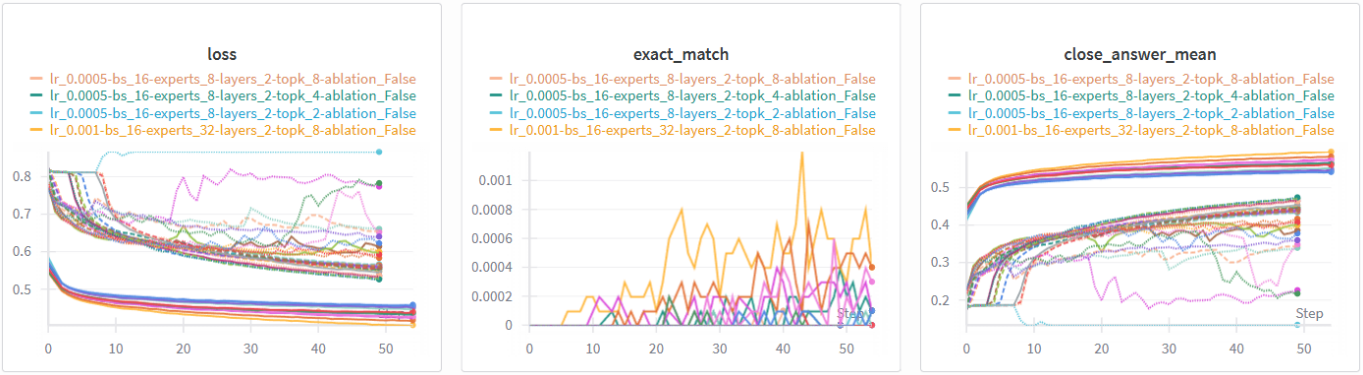


Figura 2: Resultados de sweeps en Fase 2: evolución de loss, exact_match y close_answer_mean a lo largo de las épocas para 50 configuraciones iniciales. Se observa convergencia estable y mejora en similitud semántica, con ventaja clara para variantes con MoE.

V-C. Consumo Computacional y Eficiencia

El monitoreo de métricas de GPU durante los sweeps revela un consumo altamente eficiente. La potencia de GPU se sitúa en rangos de 30-40 W, con picos menores atribuibles a operaciones de cálculo intensivo pero breves. El porcentaje de uso de potencia de GPU promedia valores bajos y estables, mientras que el límite de potencia forzado se mantiene constante alrededor de 120 W sin variaciones. La temperatura de GPU desciende ligeramente y se estabiliza en torno a 55-65 °C, indicando ausencia de sobrecarga térmica. Finalmente, la utilización de GPU promedia 50-60 %, con variaciones mínimas entre configuraciones, lo cual se corroboró en múltiples runs independientes. Estas mediciones se obtuvieron directamente de los registros de wandb, confirmando que la sparsity inducida por los mecanismos top-k (en memoria y expertos) contribuye a una ejecución eficiente, sin dependencias de hardware especializado.

Figura 3: Métricas de consumo computacional durante sweeps de Fase 2: potencia de GPU (W), uso de potencia (%), límite de potencia forzado (W), temperatura (°C) y utilización (%). Los valores estables y bajos demuestran la eficiencia del diseño sparse, adecuado para entornos de bajo recurso.

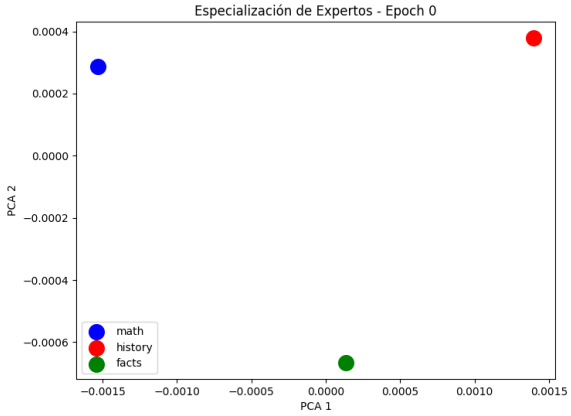
V-D. Especialización Emergente y Robustez

La visualización mediante análisis de componentes principales (PCA) de las puntuaciones medias de las puertas de routing revela un fenómeno de especialización emergente de expertos a lo largo del entrenamiento. En la época 0, los puntos correspondientes a tipos de pregunta (matemáticas, historia, hechos) aparecen superpuestos o cercanos en el espacio reducido, indicando una distribución inicial uniforme.

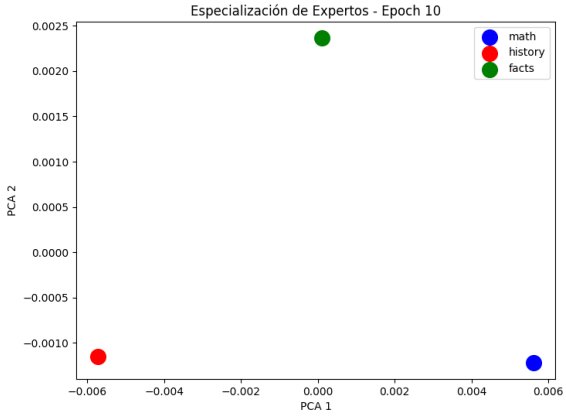
Con el avance de las épocas, los clusters se separan progresivamente: en la época 10, se observa una diferenciación inicial, que se consolida en la época 20 y alcanza claridad máxima en la época 40, con puntos bien delimitados por categoría. Este patrón se reproduce consistentemente en todas las configuraciones con MoE evaluadas, independientemente de variaciones en hiperparámetros, y no se observa en la ablación sin expertos.

Esta especialización emerge sin supervisión explícita sobre los tipos de pregunta, ya que la clasificación se utiliza únicamente para visualización posterior. La consistencia a través de múltiples sweeps y semillas de inicialización confirma que no se trata de un efecto transitorio.

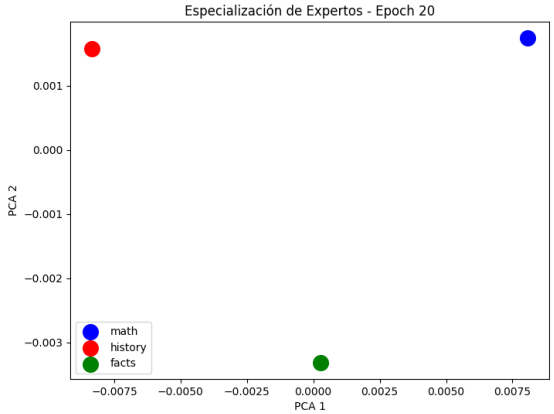
Finalmente, un hallazgo clave es la robustez del sistema completo ante recuperación imperfecta en la Fase 1. A pesar de un recall@5 bajo en evaluación (0.01), la Fase 2 genera similitudes semánticas significativas (hasta 0.6), con mejoras claras atribuidas al routing experto. Esta compensación se verifica comparando la salida de la memoria con la final del razonamiento en runs independientes.



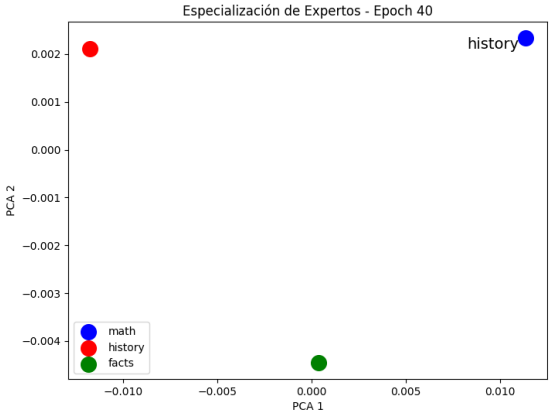
(a) Epoca 0: Distribución inicial uniforme.



(b) Epoca 10: Diferenciación inicial de clusters.



(c) Epoca 20: Consolidación de separación por categoría.



(d) Epoca 40: Clusters bien delimitados (matemáticas, historia, hechos).

Figura 4: Collage de la evolución de la especialización emergente mediante PCA de las puertas de routing a lo largo del entrenamiento. Se observa una progresión clara de superposición inicial a diferenciación semántica, confirmando el aprendizaje implícito de estructura topológica.

VI. Discusión

Los resultados obtenidos en los experimentos preliminares revelan patrones consistentes que validan la viabilidad de la arquitectura modular propuesta, a pesar de la escala reducida del prototipo. A continuación, se analiza en profundidad estos hallazgos, se identifican las limitaciones observadas y se formulan hipótesis derivadas de los datos, junto con sugerencias para experimentos futuros que permitan validarlas o refinarlas. El enfoque se centra en la interpretación bio-inspirada de los comportamientos emergentes, destacando cómo la separación estricta entre fases induce propiedades deseables en el sistema completo.

VI-A. Análisis de los Hallazgos Principales

La convergencia estable de la pérdida en la Fase 2, observada en todas las configuraciones exploradas mediante sweeps, indica que el módulo de razonamiento se optimiza de manera efectiva sobre las representaciones recuperadas de la memoria, incluso cuando estas son aproximadas o ruidosas. Esta estabilidad se mantiene independientemente de variaciones en hiperparámetros como el número de expertos o el top-k, lo que sugiere una robustez inherente al diseño modular.

La mejora progresiva en la similitud semántica media (close_answer_mean), que asciende hasta valores de 0.6en las mejores configuraciones con Mixture of Experts, demuestra que el sistema es capaz de generar respuestas coherentes a pesar de una recuperación de memoria subóptima en la Fase 1 (recall@5 aproximado de 0.01 en evaluación con ruido). Esta capacidad de compensación es particularmente notable en comparación con la ablación, donde la métrica se estanca en valores inferiores, confirmando el rol crítico del routing dinámico y la especialización de expertos en el procesamiento de información imperfecta.

Por su parte, la métrica de coincidencia exacta permanece baja en todas las variantes, lo cual es consistente con el umbral estricto de 0.9 aplicado a similitudes coseno en embeddings. Este comportamiento no indica un fallo del modelo, sino la naturaleza aproximada inherente a las representaciones vectoriales en datasets de preguntas abiertas, donde respuestas semánticamente similares pueden diferir en detalles superficiales.

En cuanto a la eficiencia computacional, las métricas de GPU registradas durante los sweeps muestran un consumo bajo y estable, con utilización promedio del 50-60 % y potencia en rangos de 30-40 W. Estos valores se mantienen constantes a través de múltiples ejecuciones, independientemente de la complejidad de la configuración (por ejemplo, número de expertos), lo que valida la efectividad de los mecanismos de sparsity inducida por top-k en ambas fases para reducir el cómputo sin comprometer la convergencia.

Finalmente, la visualización mediante PCA de las puntuaciones medias de las puertas de routing revela una evolución progresiva hacia clusters semánticos bien definidos por tipo de pregunta. Este patrón se reproduce de manera consistente en todas las configuraciones con MoE, emergiendo sin supervisión explícita sobre las categorías, lo que sugiere un aprendizaje implícito de estructura topológica en el espacio de representaciones.

VI-B. Limitaciones Observadas

A pesar de los resultados prometedores, se identifican limitaciones inherentes al alcance preliminar del prototipo. El overfitting observado en la evaluación de la Fase 1, con recall bajo bajo condiciones de ruido, se atribuye principalmente al tamaño reducido del dataset (10.000 muestras de SQuAD), que limita la diversidad de patrones almacenados y la generalización. Esta restricción se mitiga parcialmente en la Fase 2 mediante la robustez inducida por el congelamiento, pero persiste como factor que afecta las métricas finales.

La métrica de coincidencia exacta baja refleja el umbral estricto elegido, adecuado para evaluaciones precisas pero sensible a variaciones menores en embeddings, especialmente en respuestas contextuales largas. Además, el modelo opera exclusivamente en espacios de embeddings fijos (dimensión 384), lo que facilita la eficiencia pero podría limitar la expresividad en comparación con modelos token-a-token más flexibles.

El consumo computacional eficiente se logra gracias a la sparsity, pero en setups con mayor número de expertos o capas, podría requerir optimizaciones adicionales para mantener la accesibilidad en entornos de bajo recurso. Estas limitaciones no invalidan los hallazgos, sino que destacan áreas para refinamiento en iteraciones futuras.

VI-C. Hipótesis Derivadas de los Resultados

Basado en los patrones observados, se formulan las siguientes hipótesis, que explican los comportamientos emergentes y guían direcciones futuras de investigación. Estas se derivan directamente de los datos empíricos, como la mejora en similitud semántica y la evolución de los clusters en PCA.

Una hipótesis central es que, en la Fase 2, la estructura Mixture of Experts se ve obligada a extraer relaciones más profundas entre conceptos al no poseer atajos o respuestas directas provenientes de una memoria perfecta. Dado que el módulo de memoria proporciona vectores recuperados aproximados o ruidosos, los expertos deben aprender representaciones distribuidas que compensen estas imperfecciones, fomentando una especialización semántica implícita observable en las visualizaciones PCA. Esta dinámica no surge en la ablación con MLP denso, lo que sugiere que el routing dinámico es clave para forzar esta extracción relacional profunda.

Otra hipótesis es que la separación modular y el congelamiento inducen una robustez inherente al ruido en la recuperación, similar a cómo sistemas biológicos compensan recuerdos parciales mediante inferencia distribuida. Los resultados muestran que, a pesar de un recall bajo, la similitud semántica final mejora consistentemente, indicando que el razonamiento experto actúa como un mecanismo de corrección adaptativa.

Finalmente, se hipotetiza que la especialización emergente en clusters semánticos (matemáticas, historia, hechos) surge de la interacción entre el routing topológico y la clasificación implícita en el dataset, sin necesidad de supervisión explícita. La evolución progresiva de los clusters en las épocas sugiere un aprendizaje gradual de topología en el espacio de representaciones.

VI-D. Experimentos Futuros para Validar las Hipótesis

Para validar y refinar estas hipótesis, se proponen los siguientes experimentos futuros, diseñados para escalar el prototipo y controlar variables específicas.

Respecto a la hipótesis de extracción de relaciones profundas en la Fase 2, un experimento clave consistiría en escalar el dataset a conjuntos más grandes y diversos (por ejemplo, full SQuAD o combinaciones con datasets como TriviaQA o HotpotQA), midiendo si la mejora en similitud semántica se mantiene o aumenta con mayor variedad de patrones. Además, se podría introducir ruido controlado en la salida de la memoria durante el entrenamiento de la Fase 2, evaluando si esto fuerza aún más la especialización profunda y mejora la generalización, comparando con variantes sin ruido.

Para la hipótesis de robustez inducida por modularidad, experimentos futuros podrían incluir validación en benchmarks de razonamiento multi-paso (como GSM8K para matemáticas o datasets históricos), cuantificando la capacidad del sistema para

manejar consultas que requieren inferencia sobre recuperación imperfecta. Asimismo, se sugiere comparar con arquitecturas end-to-end no modulares entrenadas en los mismos datos, midiendo métricas de forgetting y robustez al ruido para confirmar la ventaja del congelamiento.

En cuanto a la hipótesis de especialización emergente, futuros trabajos podrían incorporar técnicas de análisis de interpretabilidad avanzadas (como SHAP o atención visualizada) para desglosar cómo el routing topológico contribuye a la formación de clusters. Además, escalar el número de categorías de preguntas (añadiendo tipos como çiencia.º "geografía") y evaluar la evolución de PCA en datasets más complejos permitiría verificar si la emergencia es escalable y no limitada al dataset actual.

Estos experimentos no solo validarían las hipótesis, sino que extenderían el prototipo hacia aplicaciones prácticas, como integración con generación token-a-token o despliegue en dispositivos de bajo recurso.

VII. Conclusiones

Los resultados preliminares obtenidos en este prototipo compacto validan la viabilidad de la arquitectura modular bio-inspirada propuesta. La separación estricta entre memoria asociativa (implementada mediante una red Hopfield moderna) y razonamiento dinámico (basado en Mixture of Experts) permite lograr comportamientos avanzados con menos de 15 millones de parámetros entrenables en total, ejecutados exclusivamente en recursos gratuitos como Google Colab.

Las principales conclusiones derivadas de los experimentos son las siguientes:

- La arquitectura converge de manera estable en la Fase 2, mostrando una mejora progresiva en la similitud semántica de las respuestas (close_answer_mean hasta 0.6) a pesar de una recuperación imperfecta en la Fase 1 (recall@5 aproximado de 0.01 en evaluación con ruido).
- Las configuraciones con Mixture of Experts superan consistentemente a la ablación con MLP denso en la métrica close_answer_mean, con diferencias sistemáticas de hasta 0.15 puntos observadas a través de múltiples sweeps.
- Se produce una especialización emergente clara de los expertos, visualizada mediante análisis PCA de las puntuaciones de routing, que evoluciona desde puntos superpuestos en épocas iniciales hasta clusters bien definidos por tipo de pregunta (matemáticas, historia, hechos) en épocas avanzadas.
- El consumo computacional se mantiene bajo y estable (utilización GPU promedio 50-60 %, potencia 30-40 W), confirmando la eficiencia inducida por los mecanismos de sparsity top-k en ambas fases.

Estos hallazgos demuestran que un diseño modular con entrenamiento en fases y congelamiento puede inducir robustez ante entradas imperfectas y emergencia semántica en configuraciones de recursos muy limitados. La capacidad del módulo de razonamiento para compensar las limitaciones de la memoria, junto con la especialización observada sin supervisión explícita, sugieren que principios bio-inspirados como la separación funcional y el routing condicional son prometedores para el desarrollo de sistemas de inteligencia artificial más eficientes y resilientes.

El código fuente completo, junto con los registros de wandb de todos los sweeps realizados, se encuentra disponible públicamente en un repositorio open-source para facilitar la reproducción de los experimentos y la colaboración futura. Este trabajo preliminar abre vías interesantes hacia arquitecturas que combinen memoria asociativa eficiente con razonamiento distribuido, con potencial aplicación en entornos de bajo recurso o sistemas híbridos.

Se invita a la comunidad a explorar extensiones del prototipo, como el escalado a datasets más grandes, la incorporación de generación token-a-token o la implementación de una tercera fase de decisión con loops internos, para seguir investigando las propiedades emergentes de diseños modulares bio-inspirados.

Agradecimientos

Este prototipo se desarrolló en 12 días utilizando exclusivamente recursos gratuitos como Google Colab y la asistencia del modelo Grok de xAI para depuración de código y redacción. El autor es un investigador autodidacta de 23 años sin formación universitaria formal en inteligencia artificial.

Referencias

[1] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *Advances in neural information processing systems*, vol. 29, 2016.

[2] S. Ramasinghe, D. Zoran, B. Leibe, J. Eriksen, and L. Van Gool, "Modern hopfield networks for few- and zero-shot learning," *arXiv preprint arXiv:2207.05345*, 2022.

[3] D. Krotov and J. Hopfield, "Large associative memory problem in neurobiology and machine learning," *arXiv preprint arXiv:2008.06996*, 2020.

[4] M. Rizzuto, H. Ramsauer, M. Weissensteiner, B. Schäfl, J. Lehner, M. Widrich, S. Hochreiter, and G. Klambauer, "Modern hopfield networks for clustering and outlier detection," *arXiv preprint arXiv:2401.12485*, 2024.

[5] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[6] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.