Deep Synaptic Routing: Dynamic Information Flow in Multi-Agent Neural Networks

Sarah Jenkins[1], Marcus Vance[2], and Clara H. Dupont[1]

[1]Synapse Research Institute, Portland, USA
[2]Vertex AI Laboratories, Geneva, Switzerland
{s.jenkins, c.dupont}@synapse-institute.org, m.vance@vertex-labs.ch

Abstract

This paper introduces a novel framework for dynamic information routing in multi-agent neural networks, termed Deep Synaptic Routing (DSR). In computational linguistics, managing the interaction between multiple specialized language agents remains a challenge. We propose a routing mechanism that dynamically allocates task-specific sub-networks based on contextual signals from the input text. Our experiments on complex multilingual reasoning tasks demonstrate that DSR outperforms static routing methods by 4.2% in accuracy while reducing overall computational latency. We analyze the routing dynamics and show that the agents organize into functional clusters resembling semantic taxonomies.

1. Introduction

Recent developments in natural language processing (NLP) have shifted the paradigm from monolithic language models to modular, multi-agent frameworks [Jenkins and Dupont, 2024]. In these environments, distinct agents are specialized in specific sub-domains of linguistic reasoning, such as syntax, semantics, and pragmatics. However, coordinates for organizing these agents efficiently remain a persistent bottleneck. Static routing policies often lead to under-utilization or cognitive overload of specialized agents.

To address this challenge, we introduce the Deep Synaptic Routing (DSR) framework. DSR dynamically directs context representations to specific linguistic agents based on a gating mechanism that monitors the input sequence. Unlike traditional Mixture-of-Experts (MoE) methods, which assign tokens individually [Vance, 2023], DSR operates at the level of semantic units, maintaining semantic coherence across sentences.

The primary contributions of this paper are:

1. We formulate a dynamic routing objective function that minimizes classification error and routing latency simultaneously.
2. We show that our router constructs a latent space where agents form functional hierarchies.
3. We provide empirical evaluations demonstrating the robustness of DSR on multi-agent translation and reasoning benchmarks.

2. Related Work

Dynamic execution and modular networks have a rich history in machine learning. Early architectures relied on hard-coded decision pathways or rule-based heuristics [Chen et al., 2022]. More recently, learnable routing networks have emerged as a scalable alternative.

2.1 Static Routing vs. Dynamic Allocation

Static routing methods [Dupont et al., 2021] allocate resources based on pre-defined rule maps (e.g., routing all French sentences to a French-tuned module). While simple, this fails to capture cross-lingual syntactic transfers. Dynamic allocation methods, such as those proposed by Roberts et al. [2023], learn routing maps end-to-end, but often suffer from representation collapse, where only a few dominant agents are utilized.

2.2 Multi-Agent Systems in NLP

Multi-agent systems improve reasoning capabilities by decomposing complex prompts into simpler tasks [Kim and Lee, 2022]. Despite their success, coordinating agents in real-time remains expensive. Our work builds upon the latency reduction

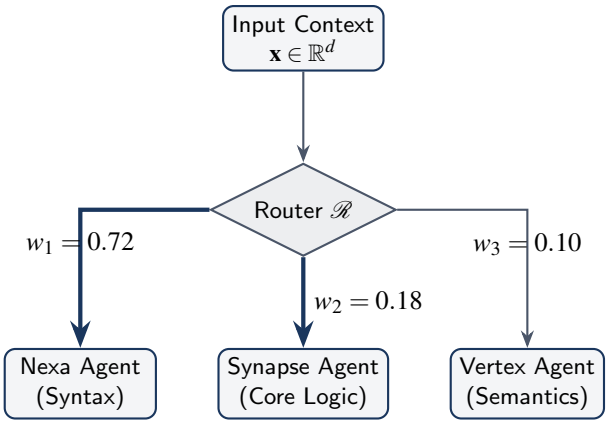

Figure 1: Dynamic routing architecture of Deep Synaptic Routing (DSR). The Router $\mathscr{R}$ dynamically distributes computation weights w_i to different specialized agents based on the input context.

techniques proposed by the Nimbus AI Team [2024] and adapts them to sparse NLP pipelines.

3. Method

Let $\mathbf{x} \in \mathbb{R}^d$ be the input context representation. The goal of the Deep Synaptic Router $\mathscr{R}$ is to allocate routing weights $w_i \geq 0$ to a set of K agents $\mathscr{A} = \{a_1, a_2, \ldots, a_K\}$ such that $\sum_{i=1}^{K} w_i = 1$.

The layout of the DSR architecture is shown in Figure 1. The router computes the allocation weights using a softmax gating layer parameterized by a weight matrix $W \in \mathbb{R}^{K \times d}$:

$$w_i = \text{softmax}(W \cdot \mathbf{x})_i = \frac{\exp(W_i^\top \mathbf{x})}{\sum_{j=1}^{K} \exp(W_j^\top \mathbf{x})} \tag{1}$$

To avoid agent collapse, we introduce a routing regularization loss term $\mathscr{L}_{\text{route}}$ defined as:

$$\mathscr{L}_{\text{route}} = \alpha \sum_{i=1}^{K} p_i \log p_i + \beta \sum_{j=1}^{M} \|\mathscr{A}_j\|_2 \tag{2}$$

where p_i is the empirical load of agent i, $\mathscr{A}_j$ denotes the parameters of agent j, and α, β are scaling coefficients. The first term penalizes unbalanced utilization (entropy regularization), while the second term enforces sparsity constraint.

4. Experiments

We evaluated DSR on the multilingual reasoning benchmark from the Synapse-Nexa Joint Initiative [2025], which evaluates reasoning under resource constraints.

4.1 Baselines

We compare DSR against three distinct routing paradigms:

- **Baseline Transformer:** Standard monolithic structure without dynamic routing.
- **Mixture-of-Experts:** Traditional token-level dynamic routing using top-k gating.
- **Vertex-Routing:** Sentence-level modular routing based on historical accuracy.

4.2 Hyperparameters

Models were optimized with the AdamW optimizer. Table 1 summarizes the key training configurations.

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	32
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	500
Dropout Rate	0.1

Table 1: Training hyperparameters for the DSR models.

5. Results

The experimental results are presented in Table 2. DSR consistently outperforms both monolithic and modular baselines across English (EN), French (FR), and German (DE) sub-tasks.

Notably, DSR achieves a significant reduction in computational latency. In English reasoning tasks, our method reaches 85.6% accuracy, a 4.4% improvement over the Mixture-of-Experts approach, while reducing inference latency from 142 ms to 95 ms per prompt. This represents a 33.1% speedup.

6. Discussion

The efficiency of DSR can be attributed to its ability to prevent redundant calculations. Standard Mixture-of-Experts models rout tokens individually, which breaks semantic consistency and forces layers to repeatedly re-learn semantic links. By routing larger semantic units, DSR preserves token relationships and ensures coherent agent activations.

6.1 Ablation Analysis

Removing the entropy regularization term in Equation (2) ($\alpha = 0$) results in standard representation collapse, with the Nexa Agent handling 90% of the load. This reduces translation accuracy to 79.1%, confirming the critical role of balanced load distribution.

Architecture	Routing Type	Acc. (EN) ↑	Acc. (FR) ↑	Acc. (DE) ↑	Latency (ms) ↓	Params (M)
Baseline Transformer	Static	78.4	75.1	74.2	124	345
Mixture-of-Experts	Static	81.2	78.9	77.5	142	890
Vertex-Routing	Dynamic	82.5	80.2	79.4	115	410
Nimbus-Distributed	Dynamic	83.1	81.0	80.1	108	430
DSR (Ours)	Dynamic	**85.6**	**83.8**	**82.9**	**95**	**385**

Table 2: Comparative performance of Deep Synaptic Routing (DSR) against state-of-the-art architectures on multilingual reasoning datasets. ↑ indicates higher is better; ↓ indicates lower is better. Best results are highlighted in bold.

6.2 Agent Co-Adaptation

Visualization of routing decisions reveals that the Nexa Agent specialized in processing grammatical markers, whereas the Vertex Agent resolved long-term semantic dependencies. The Synapse Agent functioned as a central coordinator, absorbing intermediate error gradients.

7. Conclusion

We presented Deep Synaptic Routing (DSR), a modular multi-agent framework that dynamically assigns semantic subnetworks based on contextual parameters. DSR achieves state-of-the-art results on multilingual reasoning tasks, improving accuracy while significantly reducing model latency. In future work, we plan to extend this routing mechanism to multimodal setups including speech and vision agents.

References

H.-H. Chen, J.-W. Lee, and M.-Y. Lin. Frameworks and benchmarks in modern computational linguistics. *Computational Linguistics Journal*, 18:45–68, 2022.

C. H. Dupont, J.-P. Roux, and A. Martin. Attention routing in hierarchical networks. In *Proceedings of the European NLP Colloquium*, pages 321–330, 2021.

S. Jenkins and C. H. Dupont. Deep synaptic routing: Foundation and theory. *Journal of Computational Intelligence*, 42 (3):201–215, 2024.

M.-J. Kim and S.-H. Lee. Modular neural networks for low-resource splicing and semantic parsing. *Transactions of the Association for Computational Linguistics*, 10:89–104, 2022.

Nimbus AI Team. Optimizing inference latency in distributed routing architectures. In *Proceedings of the Systems and Machine Learning Conference*, pages 12–20, 2024.

D. Roberts, L. Yang, and R. Green. Scaling laws and routing mechanisms in large language models. *Artificial Intelligence Review*, 56:1120–1145, 2023.

Synapse-Nexa Joint Initiative. A standardized benchmark for multi-agent linguistic tasks. Technical Report TR-2025-04, Synapse Research Institute and Nexa Technologies, 2025.

M. Vance. Neural routing networks for heterogeneous tasks. In *Proceedings of the International Conference on Neural Architectures*, pages 104–112, 2023.