Arquitetura Neuro-Simbólica Híbrida Baseada em Transformer, NEF/Spaun e Plasticidade Online Local: Derivação Matemática de um Modelo Cognitivo Escalável

Natã Fernandes

25 de abril de 2026

Resumo

Este trabalho propõe uma arquitetura neuro-simbólica híbrida que integra mecanismos de atenção do tipo Transformer com a estrutura neurodinâmica do *Neural Engineering Framework* (NEF), empregado no projeto Spaun, incorporando ainda memória persistente, roteamento modular inspirado nos gânglios da base e plasticidade sináptica local online. O objetivo central é investigar uma alternativa ao paradigma de aprendizado massivo predominantemente offline que caracteriza grandes modelos contemporâneos. A arquitetura proposta, denominada *Hybrid Attention Transformer-Spaun* (HATS), combina atenção auto-regressiva contextual, codificação neural vetorial contínua, memória episódica e semântica dinâmica, roteamento modular e aprendizado online local. Neste texto, derivamos formalmente cada componente, explicitamos suas relações matemáticas e apresentamos uma formulação unificada para um sistema cognitivo escalável, interpretável e compatível com atualização incremental.

1 Introdução

Modelos Transformer modernos exibem capacidades emergentes expressivas, porém sua escalabilidade prática está associada a custos computacionais e energéticos elevados. Em primeira aproximação, pode-se modelar esse custo pela relação

$$C \propto N_p \cdot D \cdot T, \tag{1}$$

em que N_p representa o número de parâmetros, D denota o volume de dados de treinamento e T corresponde ao tempo computacional necessário para a otimização global do sistema.

Em contraste, o aprendizado biológico parece obedecer a uma dinâmica qualitativamente distinta, mais dependente da relevância da experiência e da plasticidade local do que de ciclos extensivos de reprocessamento global. Em forma esquemática, tal comportamento pode ser expresso como

$$L \propto E \cdot P, \tag{2}$$

onde E representa experiências relevantes e P indica plasticidade neural.

O problema abordado neste artigo consiste em derivar matematicamente uma arquitetura capaz de aproximar a segunda dinâmica sem abandonar os avanços computacionais trazidos pelos mecanismos de atenção. A hipótese central é que uma integração entre atenção contextual, representação neural contínua, memória dinâmica, roteamento modular e plasticidade local pode oferecer um caminho promissor para sistemas de inteligência artificial mais incrementais, energeticamente eficientes e cognitivamente estruturados.

2 Fundamentos Matemáticos do Transformer

2.1 Embeddings

Seja uma sequência de entrada

$$X = (x_1, x_2, \ldots, x_n), \tag{3}$$

na qual cada x_i representa a codificação simbólica do i-ésimo elemento da sequência. A etapa inicial do Transformer consiste em projetar cada token para um espaço vetorial contínuo por meio de uma matriz de embeddings:

$$E_i = W_e x_i. \tag{4}$$

Para incorporar informação de ordem, adiciona-se uma codificação posicional determinística. No formalismo senoidal clássico, obtém-se

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \tag{5}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right). \tag{6}$$

Assim, a representação final fornecida ao bloco atencional é

$$Z_i = E_i + PE_i. \tag{7}$$

2.2 Autoatenção

Uma vez definida a matriz de entrada Z, constroem-se as projeções lineares de consulta, chave e valor:

$$Q = ZW_Q, \tag{8}$$

$$K = ZW_K, \tag{9}$$

$$V = ZW_V. \tag{10}$$

O produto escalar bruto entre consultas e chaves é dado por

$$S = QK^T. \tag{11}$$

Após normalização pela dimensão das chaves, obtém-se

$$\tilde{S} = \frac{QK^T}{\sqrt{d_k}}. \tag{12}$$

A matriz de pesos atencionais é então calculada por

$$A = \text{softmax}(\tilde{S}), \tag{13}$$

e a saída do mecanismo de atenção assume a forma

$$H = AV. \tag{14}$$

Em notação compacta, o operador de atenção pode ser escrito como

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \tag{15}$$

3 Estrutura Matemática do NEF e do Spaun

3.1 Codificação neural

No *Neural Engineering Framework*, um estado contínuo é representado em um espaço vetorial de dimensão finita:

$$x \in \mathbb{R}^d. \tag{16}$$

A atividade do neurônio i é modelada por

$$a_i(x) = G_i \left(\alpha_i \langle e_i, x \rangle + J_i^{\text{bias}} \right), \tag{17}$$

onde G_i é a função de ativação do neurônio, e_i é seu vetor codificador, α_i é o ganho e J_i^{bias} é a corrente de polarização.

3.2 Decodificação

A reconstrução aproximada do estado contínuo a partir da população neural é dada por

$$\hat{x} = \sum_i a_i(x) d_i, \tag{18}$$

em que d_i denota o vetor decodificador associado ao neurônio i. O erro de reconstrução pode ser expresso como

$$E = \|x - \hat{x}\|^2. \tag{19}$$

Ao minimizar o erro quadrático regularizado em relação à matriz de decodificação, obtém-se a solução

$$D = (A^T A + \lambda I)^{-1} A^T X, \tag{20}$$

na qual A representa a matriz de atividades neurais, X representa os vetores-alvo e λI é o termo de regularização.

4 Dinâmica Neural Temporal

O projeto Spaun opera sobre sistemas dinâmicos contínuos. Em sua forma linear mais elementar, a dinâmica do estado pode ser escrita como

$$\dot{x} = Ax + Bu, \tag{21}$$

com A descrevendo a dinâmica interna do sistema e B acoplando a entrada u ao estado.

No formalismo neural temporal, a dinâmica também pode ser escrita como

$$\tau \dot{x}(t) = -x(t) + f(x, u), \tag{22}$$

onde τ é a constante de tempo do sistema e f representa a transformação efetivamente implementada pela população neural.

Sob discretização temporal explícita com passo Δt, obtém-se

$$x_{t+1} = x_t + \Delta t(Ax_t + Bu_t). \tag{23}$$

Essa passagem ao regime discreto permite integrar o formalismo do NEF com arquiteturas computacionais sequenciais e iterativas, como os Transformers.

5 Integração Transformer + NEF

Seja H_t a saída do bloco Transformer no instante t. Projetamos essa saída para um espaço neural contínuo por meio de uma transformação linear:

$$z_t = W_h H_t. \tag{24}$$

O vetor z_t é então codificado na população neural segundo

$$a_i(z_t) = G_i \left(\alpha_i \langle e_i, z_t \rangle + b_i \right), \tag{25}$$

com b_i representando um termo de polarização local. A reconstrução vetorial obtida a partir da atividade da população é

$$\hat{z}_t = \sum_i a_i(z_t) d_i. \tag{26}$$

Inserindo essa reconstrução numa dinâmica contínua do tipo NEF, temos

$$\dot{z} = A\hat{z} + Bu, \tag{27}$$

ou, de forma expandida,

$$\dot{z} = A \left(\sum_i a_i(z) d_i \right) + Bu. \tag{28}$$

Essa equação representa a primeira fusão formal entre atenção contextual simbólica e dinâmica neural contínua distribuída.

6 Memória Persistente Dinâmica

Para introduzir persistência temporal e retenção adaptativa de estados internos, definimos uma matriz de memória

$$M_t \in \mathbb{R}^{m \times d}. \tag{29}$$

A regra de atualização proposta é

$$M_{t+1} = \gamma M_t + \eta z_t z_t^T, \tag{30}$$

em que γ é o fator de retenção e η é a taxa de aprendizagem associada à escrita na memória.

Por expansão recursiva, a memória total acumulada até o instante t assume a forma

$$M_t = \sum_{k=0}^{t} \gamma^{t-k} \eta z_k z_k^T. \tag{31}$$

Essa expressão explicita que o sistema mantém um traço exponencialmente decrescente das representações anteriores, o que permite interpretar M_t como um acoplamento entre memória episódica e memória semântica emergente.

7 Roteamento Modular Inspirado nos Gânglios da Base

Seja um conjunto de módulos funcionais

$$\mathcal{M} = \{M_1, M_2, \ldots, M_n\}. \tag{32}$$

O roteamento entre esses módulos é modelado por um mecanismo de *gating* suave, cujos pesos são dados por

$$g_i = \frac{e^{u_i}}{\sum_j e^{u_j}}, \tag{33}$$

onde u_i representa a utilidade ou saliência interna do módulo M_i. A saída global do sistema é então escrita como

$$Y = \sum_i g_i M_i(x). \tag{34}$$

Essa formulação confere seletividade funcional à arquitetura, permitindo que apenas subconjuntos relevantes de processamento sejam ativados em cada contexto computacional.

8 Plasticidade Local Online

Em vez de depender exclusivamente de retropropagação global, a arquitetura proposta incorpora atualização local de parâmetros segundo regras biologicamente inspiradas.

8.1 Regra Hebbiana

A forma mais simples de plasticidade local pode ser escrita como

$$\Delta w_{ij} = \eta x_i y_j, \tag{35}$$

onde x_i representa a atividade pré-sináptica, y_j representa a atividade pós-sináptica e η é a taxa local de aprendizagem.

8.2 Regra temporal do tipo STDP

Para capturar dependência temporal entre disparos, considera-se a regra

$$\Delta w = \begin{cases} A_+ e^{-\Delta t/\tau_+}, & \Delta t > 0, \\ -A_- e^{\Delta t/\tau_-}, & \Delta t < 0, \end{cases} \tag{36}$$

em que Δt denota a diferença temporal entre eventos pré- e pós-sinápticos, A_+ e A_- são amplitudes de potenciação e depressão, e τ_+ e τ_- são constantes temporais características.

9 Formulação Completa do Modelo HATS

Em nível estrutural, a arquitetura pode ser descrita pela cadeia

$$X \to \text{Transformer} \to \text{NEF} \to \text{Memória} \to \text{Gating} \to \text{Output}. \tag{37}$$

De forma abstrata, a composição funcional total do sistema é

$$Y_t = \mathcal{G}\Big(\mathcal{M}(\mathcal{N}(T(X_t)))\Big), \tag{38}$$

onde T denota o bloco Transformer, $\mathcal{N}$ a transformação neurodinâmica do tipo NEF, $\mathcal{M}$ o operador de memória persistente e $\mathcal{G}$ o mecanismo de roteamento modular.

Expandindo a expressão acima, obtém-se a formulação

$$Y_t = \sum_i g_i \, M_i \left(\sum_j a_j \left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \right) V \right) d_j \right). \tag{39}$$

Essa expressão sintetiza a proposta central do modelo HATS: unir atenção contextual, representação neural contínua, memória dinâmica, modulação funcional e plasticidade local dentro de uma única arquitetura.

10 Complexidade Computacional

Para um Transformer padrão, o custo dominante da autoatenção é da ordem de

$$\mathcal{O}(n^2 d), \tag{40}$$

com n representando o comprimento da sequência e d a dimensão da representação.

Na arquitetura proposta, o custo total pode ser aproximado por

$$\mathcal{O}(n^2 d) + \mathcal{O}(kd) + \mathcal{O}(md), \tag{41}$$

em que k representa o número de neurônios efetivamente ativos e m representa a fração de memória dinamicamente acessada.

Como a arquitetura assume ativação esparsa e roteamento seletivo, considera-se

$$k \ll N, \tag{42}$$

onde N representa o número total de unidades disponíveis. Em princípio, isso pode reduzir de forma significativa o custo energético e a necessidade de retreinamento global do sistema.

11 Propriedades Emergentes Esperadas

Do ponto de vista funcional, a arquitetura HATS pretende oferecer as seguintes propriedades:

- aprendizado contínuo e incremental;
- memória persistente com atualização dinâmica;
- raciocínio modular por seleção contextual de módulos;
- menor dependência de retreinamento massivo;
- maior interpretabilidade em comparação com arquiteturas monolíticas puramente paramétricas.

Essas propriedades decorrem da combinação entre mecanismos de atenção, representação distribuída, memória estruturada e plasticidade local.

12 Limites Teóricos e Condições de Estabilidade

A estabilidade da dinâmica interna exige restrições sobre os operadores de evolução do sistema. Uma condição suficiente para estabilidade assintótica do componente linear é

$$\rho(A) < 1, \tag{43}$$

onde $\rho(A)$ denota o raio espectral da matriz A.

Para a memória persistente, a convergência da soma geométrica requer

$$0 < \gamma < 1. \tag{44}$$

Já a plasticidade local deve obedecer a um limite superior de intensidade para evitar divergência dinâmica:

$$\eta < \eta_{\text{crit}}. \tag{45}$$

Tais condições definem um regime de operação no qual a arquitetura permanece atualizável sem perda de estabilidade global.

13 Conclusão

Demonstramos matematicamente uma nova classe arquitetural híbrida que integra atenção do tipo Transformer, representação neural contínua baseada no NEF, memória dinâmica persistente, plasticidade local online e roteamento cognitivo modular. A arquitetura HATS foi formulada como tentativa de superar a dependência exclusiva do paradigma de escalonamento massivo, introduzindo uma via alternativa fundamentada em atualização incremental, modularidade funcional e aprendizagem local.

A hipótese central desenvolvida neste trabalho é que sistemas futuros de inteligência artificial podem beneficiar-se de uma transição parcial do modelo puramente paramétrico e offline para um regime mais orgânico, adaptativo e cognitivamente estruturado. Nesse sentido, o valor da proposta não reside apenas em sua formulação conceitual, mas na possibilidade de oferecer uma base matemática para arquiteturas capazes de unir desempenho contextual, persistência de memória, plasticidade contínua e interpretabilidade funcional.

Referências

[1] Eliasmith, C. et al. *How to Build a Brain*. 2012.

[2] Vaswani, A. et al. *Attention Is All You Need*. 2017.

[3] Friston, K. *The Free Energy Principle*. 2010.

[4] Hinton, G. *Forward-Forward Algorithm*. 2022.