

Q1: How Sparse Is the Explanation?

Claude and MJC

11 February 2026

Abstract

We answer the first question from “Toward Total Interpretation of a Small RNN”: for each prediction of the sat-rnn (128 hidden units, 0.079 bpc on 1024 bytes), how many of the ~ 3048 patterns in the isomorphic UM participate in the backward attribution chain above a given threshold? We compute the full backward trace for all 1024 positions and report the distribution of active patterns per prediction.

1 Setup

We use the sat-rnn and its isomorphic UM u_{iso} as defined in the companion paper. The model has three weight matrices giving three classes of patterns:

- W_x patterns: input byte $\rightarrow h_j^\pm$ (up to $256 \times 128 = 32,768$)
- W_h patterns: $h_j^\pm \rightarrow h_k^\pm$ (up to $128^2 = 16,384$)
- W_y patterns: $h_j^\pm \rightarrow$ output byte (up to $128 \times 256 = 32,768$)

With significance threshold $\epsilon > 0$, approximately 3048 patterns survive. Each pattern has a strength (the absolute weight value) and a sign (excitatory or inhibitory).

2 Method

For each position $t = 0, \dots, 1023$, predicting $y = x_{t+1}$:

1. Compute the output gradient $g_t \in \mathbb{R}^{128}$ (Definition 4 of the companion paper).
2. For each offset $d = 1, \dots, D_{\max}$, compute the backward gradient $g_{t,d}$ via the Jacobian chain (Definition 5).
3. For each W_y pattern (j, y) : the pattern’s attribution is $|(g_t)_j \cdot \mathbf{1}[h_j \text{ has correct sign}]|$.
4. For each W_x pattern (x_{t-d}, j) at offset d : the pattern’s attribution is $|\alpha_j(t, d)| = |W_x[j, x_{t-d}] \cdot [g_{t,d}]_j|$.
5. For each W_h pattern (j, k) : the pattern’s attribution at offset d is $|(1 - h_j(t-d)^2) \cdot W_h[k, j] \cdot [g_{t,d}]_j|$. A W_h pattern may be active at multiple offsets; we take the maximum.

A pattern is *active* for position t if its attribution exceeds threshold τ . We sweep τ over several orders of magnitude.

2.1 Counting

For each position t and threshold τ , we report:

- $n_x(t, \tau)$: number of active W_x patterns
- $n_h(t, \tau)$: number of active W_h patterns
- $n_y(t, \tau)$: number of active W_y patterns
- $n(t, \tau) = n_x + n_h + n_y$: total active patterns

3 Results

The model has 44,794 patterns with $|w| > 0.01$ (5,371 W_x , 14,245 W_h , 25,178 W_y). The model achieves 0.079 bpc.

3.1 Sparsity distribution

| Threshold τ | Mean n | Median n | Min | Max | $n/44794$ |
|------------------|----------|------------|-----|-------|-----------|
| 10^{-4} | 9807 | 10283 | 0 | 19850 | 0.219 |
| 10^{-3} | 4357 | 1664 | 0 | 19352 | 0.097 |
| 10^{-2} | 1166 | 15 | 0 | 17710 | 0.026 |
| 10^{-1} | 157 | 0 | 0 | 11012 | 0.004 |
| 1.0 | 8 | 0 | 0 | 2127 | 0.000 |

The distribution is highly skewed: mean \gg median at every threshold. Most positions need very few patterns; a small number of positions (those with high bpc, i.e. surprising predictions) activate thousands. At $\tau = 0.01$, the median position uses only 15 patterns.

3.2 Breakdown by pattern class

| Threshold τ | Mean n_x | Mean n_h | Mean n_y |
|------------------|------------|------------|------------|
| 10^{-3} | 481 | 3834 | 42 |
| 10^{-2} | 136 | 1018 | 12 |
| 10^{-1} | 22 | 134 | 2 |

W_h patterns dominate at every threshold, accounting for $\sim 87\%$ of active patterns. This is the recurrent signal: the backward chain flows primarily through W_h connections. W_y patterns are the fewest (12 at $\tau = 0.01$), meaning the output layer is sparse—only a handful of neurons contribute meaningfully to each prediction.

3.3 Never-active patterns

At $\tau = 0.01$, 57,335 of 81,920 total patterns (70%) are never active at any position. The breakdown:

- W_x : 28,635/32,768 never active (87%)—most input bytes never occur, so most W_x patterns are irrelevant.
- W_h : 735/16,384 never active (4.5%)—nearly all recurrent connections matter somewhere.
- W_y : 27,965/32,768 never active (85%)—most output bytes are never the target.

3.4 Depth profile

Attribution mass does not decay monotonically with offset:

| Offset d | Mean mass | Fraction of $d=0$ |
|------------|-----------|-------------------|
| 0 | 0.757 | 1.000 |
| 1 | 0.176 | 0.233 |
| 2 | 0.181 | 0.239 |
| 4 | 0.243 | 0.321 |
| 8 | 0.342 | 0.451 |
| 12 | 0.406 | 0.536 |
| 20 | 0.729 | 0.963 |
| 21 | 0.827 | 1.093 |
| 30 | 0.421 | 0.556 |
| 40 | 0.571 | 0.754 |
| 50 | 0.421 | 0.557 |

The gradient does not vanish. Mass grows from $d=1$ to a peak at $d \approx 20\text{--}21$ (exceeding $d=0$), then oscillates around $0.5\text{--}0.7 \times$ the $d=0$ value out to $d=50$. The RNN mixes information into a carrier arising from its recurrent dynamics.

4 Discussion

The answer to Q1 is: **the explanation is very sparse for typical predictions but heavy-tailed.** The median position at $\tau = 0.01$ uses only 15 patterns out of 44,794. But the mean is 1,166, pulled up by a minority of positions with large attribution counts.

W_h dominates: the recurrent patterns are the backbone of the explanation. Nearly all (95.5%) of the W_h patterns are active at some position, confirming that the 128-neuron recurrent core is fully utilized.

The non-monotonic depth profile shows that the RNN sustains information flow well beyond the first few timesteps, with a peak at $d \approx 20$. This is consistent with the skip- k -gram finding that offset 20 was selected third in the greedy MI ordering $[1, 8, 20, \dots]$.

Reproducibility

Tool: `q1_sparsity.c` in `docs/archive/20260211/`. Model: `sat_model.bin` from `archive/20260209/`. Data: first 1024 bytes of `enwik9`.