SN Visibility for the Doubled-E Universal Model

Claude (with MJC)

2026-02-02

1 Introduction

This document describes the SN visibility implementation for the Hutter RNN, building on the RNN-to-UM mapping established in `rnn-um-mapping.tex` (archive 20260131).

The goal is to make the Universal Model visible in human-readable SN format: events with names, patterns with strengths derived from the isomorphic mapping.

2 Background

The doubled-E mapping (Section 6 of `rnn-um-mapping.tex`) establishes an exact isomorphism between the RNN and a Universal Model:

- Each hidden neuron j becomes a binary event space $\{h_j^+, h_j^-\}$
- Support values: $t(h_j^+) = 2 \cdot \max(0, \text{pre}_h[j])$, $t(h_j^-) = 2 \cdot \max(0, -\text{pre}_h[j])$
- Binary softmax replaces tanh, giving exact bpc equivalence (0.00% difference)

3 Event Structure

The model has 768 events, numbered 1–768:

Range	Count	Description
1–256	256	Input bytes: "The input is 'a'."
257–512	256	Hidden neurons: "h1+" "h1-" ... "h128+" "h128-"
513–768	256	Output bytes: "The output is 'a'."

Event naming convention:

- Use "input" and "output", not "previous/next character"
- Printable bytes shown as character: "The input is 'e'."
- Non-printable shown as hex: "The input is 0x0A."
- Hidden events use doubled-E notation: hN+ (activated), hN- (inhibited)

4 Pattern Extraction

Pattern strength equals the contribution to support in the UM forward pass:

$$\text{strength} = \lfloor 2 \cdot |w| + 0.5 \rfloor$$

This follows directly from the doubled-E mapping where weights contribute $2|w|$ to support values.

Patterns with strength < 1 are omitted—they contribute negligibly to the model's predictions.

4.1 Pattern Types

Input → **Hidden** (from W_{xh}):

- $w > 0$: pattern from input event to h_j^+
- $w < 0$: pattern from input event to h_j^-

Hidden → **Output** (from W_{hy}):

- $w > 0$: pattern from h_j^+ to output event
- $w < 0$: pattern from h_j^- to output event

Note: Hidden-to-hidden patterns (W_{hh}) are not currently exported. They represent recurrent connections and would significantly increase pattern count.

5 Results

For the trained model (`model.bin`, 5.69 bpc on enwik9):

Metric	Value
Total events	768
Total patterns	302
Input events with patterns	17/256
Hidden events with patterns	68/256
Output events with patterns	217/256

Most input bytes have no significant patterns (strength ≥ 1). The model's input-to-hidden weights are sparse—only a few bytes (space, newline, common letters) have weights ≥ 0.5.

The strongest pattern is space → h3- with strength 8 (weight ≈ 4.0).

6 SN Viewer

The interactive viewer (`sn-view.html`) provides:

- Three-panel layout: fan-in, events, fan-out
- Events with patterns shown in black; others grayed out
- Per-section statistics (e.g., "17/256 with patterns")
- Click any event to see its incoming and outgoing patterns

7 Implementation

New CLI mode added to `hutter.c`:

```
./hutter sn [model] [output_dir]
```

Outputs:

- `events.sn`: 768 events with names and 0 support
- `patterns.sn`: patterns with strength = $2|w|$

8 Interpretation

The sparsity of significant patterns reflects the model's learned structure:

- Most byte transitions are handled by small, distributed weights
- Only a few input bytes (space, common letters) have strong dedicated patterns
- The hidden layer concentrates information—68 of 256 hidden events participate in significant patterns
- The output layer is dense—217 of 256 output events receive significant patterns

This is consistent with the model learning character-level statistics where most of the "work" happens in the hidden-to-output mapping, while input-to-hidden is more uniform.

9 Relation to ES Discovery

This visibility is a prerequisite for ES discovery (the main experimental problem). With patterns now visible:

1. We can identify which hidden neurons have similar input patterns (ES candidates)
2. We can see which outputs are predicted by similar hidden configurations
3. We can propose coarser ESs and test whether bpc is preserved

The doubled-E representation is mechanical, not natural. The next step is finding semantically meaningful ESs that collapse the 128 binary ESs into fewer, interpretable ones.

10 Files

`events.sn`	Event declarations
`patterns.sn`	Pattern triples (source, destination, strength)
`sn-view.html`	Interactive viewer
`model.bin`	Trained RNN weights