Interpretation of the Isomorphic Universal Model
Tock 1 Analysis

Claude

2026-02-04

1 Introduction

This paper interprets the isomorphic Universal Model (UM) derived from an Elman RNN trained on enwik9 (5.69 bpc). The UM operates entirely in the domain of positive patterns, positive events, and positive integer support values.

The doubled-E mapping provides an exact isomorphism: the UM achieves 0.00% bpc difference from the RNN. Our task is to find the natural event spaces and named patterns within this representation, preparing for the next tick of training.

2 Model Structure

2.1 Events

The model has 768 events organized into three layers:

Layer	Count	Events	Example
Input	256	"The input is X."	"The input is 'e'."
Hidden	256	Doubled-E: h_j^+, h_j^-	"h2+", "h2-"
Output	256	"The output is X."	"The output is 'e'."

Each hidden neuron j contributes a binary event space $\{h_j^+, h_j^-\}$. This is the mechanical doubled-E representation; natural ESs may be coarser.

2.2 Patterns

Patterns connect events with positive integer strengths. For weight w:

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

Patterns with strength ≥ 1 are significant (weight ≥ 0.5).

Type	Count	Example
Input → Hidden	49	"The input is ' '." "h2-" 8.
Hidden → Output	253	"h2+" "The output is 'e'." 2.
Hidden → Hidden	(not exported)	"h2+" "h35+" 1.

Total extracted patterns: 302 (input→hidden and hidden→output only).

3 Discovered Event Spaces

3.1 ES1: Word Boundary

The binary ES $\{h_2^+, h_2^-\}$ functions as a word boundary detector.

Input patterns to ES1:

```
"The input is ' '." "h2-" 8.
"The input is 0x0A." "h2-" 1.
"The input is 't'." "h2+" 1.
"The input is 'r'." "h2+" 1.
"The input is 'e'." "h2+" 1.
... (all letters → h2+ with strength 1)
```

The space input sends support 8 to h_2^-. Letters send support 1 to h_2^+. The 8:1 ratio creates a binary switch via softmax within the ES.

Output patterns from ES1:

```
"h2+" "The output is 'e'." 2.
"h2+" "The output is 'n'." 2.
"h2+" "The output is 'l'." 1.
"h2+" "The output is 'r'." 1.
```

When h_2^+ wins (word-internal), common letters are predicted.

Interpretation:

- h_2^+ = "word-internal position"
- h_2^- = "word boundary"

3.2 ES2: Syllable Momentum

The binary ES $\{h_{35}^+, h_{35}^-\}$ tracks syllable structure.

Input patterns:

```
"The input is ' '." "h35-" 2.
"The input is 'e'." "h35+" 1.
```

Vowels send weak support to h_{35}^+. Space sends support 2 to h_{35}^-.

Recurrent pattern (critical):

```
"h2+" "h35+" 1.   (weight = 0.612)
```

When h_2^+ is active (word-internal), it drives h_{35}^+. This creates syllable momentum within words.

Self-connection:

$$W_{hh}[35, 35] = -0.184$$

The negative self-connection means h_{35}^+ decays over time. In UM terms: h_{35}^+ sends support to h_{35}^-, causing gradual shift.

Interpretation:

- h_{35}^+ = "syllable momentum" (high early in word)
- h_{35}^- = "momentum decayed" (high late in word)

This explains why short words end with high h_{35}^+ support and long words end with high h_{35}^- support.

3.3 Natural ES Discovery: Correlation Analysis

Correlation analysis reveals massive redundancy in the 128 hidden neurons. Many pairs have $r = 1.000$ or $r = -1.000$, meaning they're functionally identical or opposite.

Major clusters (neurons with $r > 0.9$):

- Cluster A (13 neurons): h12, h20, h39, h40, h41, h58, h59, h64, h66, h83, h85, h112, h115
- Cluster B (11 neurons): h1, h4, h16, h21, h49, h74, h95, h98, h101, h121, h124
- Cluster C (11 neurons): h5, h8, h28, h31, h38, h44, h65, h68, h77, h105, h118
- Cluster D (9 neurons): h11, h17, h22, h24, h26, h53, h81, h87, h117
- Cluster E (9 neurons): h15, h37, h79, h89, h92, h93, h100, h102, h116
- Additional smaller clusters...

Implication: The 128 binary ESs (256 events) can collapse to ∼18 natural ESs. Each cluster acts as a single binary ES—all members flip together.

3.4 Saturated Neurons

Always ON (mean > 0.9): h0, h10, h19, h34, h48, h103, h109

These neurons saturate at $h^+ \approx 255$ support. They represent constant context ("processing text") rather than varying information.

Always OFF (mean < -0.9): h6, h23, h25, h32, h90, h91

These saturate at $h^- \approx 255$. Combined with the always-ON neurons, about 13 neurons (10%) carry no dynamic information.

3.5 Active Binary ESs

The remaining neurons form active binary ESs. Key ones:

- ES(h_2): word boundary, h^+ dominates 86.4%
- ES(h_3): (role unclear), h^+ dominates 33.3%
- ES(h_{35}): syllable momentum, h^+ dominates 63.6%
- ES(h_{62}): (correlated with word-internal), h^+ dominates 66.7%

4 Pattern Inventory

4.1 Strong Input Patterns

Patterns from input events to hidden events with strength ≥ 2:

Input Event	Hidden Event	Strength
"The input is ' '."	h2-	8
"The input is ' '."	h62-	8
"The input is ' '."	h72-	6
"The input is ' '."	h3-	8
"The input is ' '."	h115-	3
"The input is ' '."	h35-	2

Space is the dominant input—it resets multiple hidden states.

4.2 Strong Output Patterns

The hidden→output patterns are more distributed. Top patterns:

Hidden Event	Output Event	Strength
h2+	"The output is 'e'."	2
h2+	"The output is 'n'."	2
(many with strength 1)		

Most output patterns have strength 1, reflecting distributed encoding.

5 Redistributing the Negative Signal

The doubled-E representation handles negative weights mechanically: negative weight to h_j becomes positive pattern to h_j^-.

For natural ES discovery, we should find groups where h_j^- and h_k^+ (for different j, k) form a single ES.

5.1 Candidate: h2 and h62

Both h_2 and h_{62} receive strong patterns from space:

```
"The input is ' '." "h2-" 8.
"The input is ' '." "h62-" 8.
```

And both receive weak patterns from letters.

Hypothesis: $\{h_2^+, h_2^-, h_{62}^+, h_{62}^-\}$ might collapse to a smaller ES if h_2 and h_{62} are redundant.

To test: check if h_2 and h_{62} activations are correlated.

5.2 Analysis Needed

To properly redistribute:

1. Find hidden neurons with correlated activations
2. Identify which h_j^- events co-occur with which h_k^+ events
3. Propose coarser ESs that preserve bpc

6 Interpretation Coverage

6.1 Events Interpreted

Layer	Total	Interpreted	Coverage
Input	256	17	6.6%
Hidden	256	4 (h2, h35, h62, h3)	1.6%
Output	256	217 (receive patterns)	84.8%
Total	768	238	31.0%

"Interpreted" for hidden means we have a semantic name. Most hidden neurons remain unnamed.

6.2 Patterns Interpreted

Of the 302 significant patterns:

- 6 space→hidden patterns with strength ≥ 2: understood as "reset" patterns
- 49 total input→hidden: mostly strength 1, representing character identity
- 253 hidden→output: distributed character prediction

Interpretation: About 10 patterns have clear semantic meaning (word boundary, syllable tracking). The rest are mechanical character-level statistics without higher-level interpretation.

6.3 BPC Attribution

The model achieves 5.69 bpc. How much is explained by interpreted structure?

ES1 (word boundary) contributes to predicting:

- Space after words
- Common word-initial letters

Rough estimate: word boundaries represent ∼15% of predictions (average word length ∼5-6 chars). But ES1 doesn't fully determine these predictions—it's one signal among many.

Conservative estimate: Interpreted patterns explain $< 5\%$ of compression. The bulk of the model's 5.69 bpc performance comes from uninterpreted character-level statistics in the hidden→output weights.

7 Preparation for Next Tick

7.1 What We Have

1. ES1 (word boundary): clean binary ES, 99.6% separation
2. ES2 (syllable momentum): explains word-length encoding
3. SN pattern format for the model
4. Understanding that words are NOT explicitly encoded

7.2 What Remains

1. Find natural ESs beyond the mechanical doubled-E
2. Interpret more hidden neurons (currently 4/128)
3. Understand the hidden→hidden recurrent structure
4. Determine which patterns can be injected for the tick

7.3 Injection Candidates

For the next tick, we could inject:

- Strengthen the word-boundary pattern (currently strength 8)
- Add patterns for common word endings
- Add patterns for frequent bigrams/trigrams

The tick should write patterns that encode lexical knowledge the model hasn't learned from data alone.

8 Conclusion

The isomorphic UM interpretation reveals:

1. A clear word-boundary ES (h_2) driven by space patterns
2. A syllable momentum ES (h_{35}) with temporal decay
3. Distributed character-level prediction in output patterns
4. No explicit word identity encoding

Coverage: $\sim$31% of events touched, $< 5\%$ of bpc explained. The model is primarily a character-level statistical predictor with word-boundary structure emerging from the space pattern.

For the next tick: inject lexical patterns to bootstrap word knowledge.