Synthesis: Eight Days of RNN Interpretability Observations and Foundations for Continued Analysis

Claude

2026-02-06

Abstract

This paper synthesizes eight days of intensive research on interpreting an Elman RNN trained on enwik9 for text compression. We collect the key observations, validated findings, refuted hypotheses, and theoretical foundations that will guide continued analysis. The work establishes a tick-tock methodology for RNN-UM (Universal Model) bidirectional translation, discovers interpretable Event Spaces for word boundaries and syllable structure, and develops a unified theoretical framework connecting Bayesian inference, thermodynamics, and neural computation.

Contents

1 Model and Performance Baseline

1.1 Architecture

The model is an Elman RNN with architecture 256→128→256:

- Input: 256 one-hot bytes
- Hidden: 128 tanh units with recurrent connections
- Output: 256 softmax over next byte

1.2 Performance

Metric	Value
Training data	enwik9 (1B bytes Wikipedia XML)
Performance	5.69 bpc
Random baseline	∼8.0 bpc
Information captured	∼2.3 bits/char
State of art	∼1.1 bpc

1.3 The Isomorphic Universal Model

We established an exact isomorphism between the RNN and a Universal Model:

- **Doubled-E representation**: Each hidden neuron j becomes binary ES $\{h_j^+, h_j^-\}$
- **Support mapping**: $t(h_j^+) = 2\max(0, \text{pre}_h[j])$, $t(h_j^-) = 2\max(0, -\text{pre}_h[j])$
- **Pattern strength**: strength $= \lfloor 2|w| + 0.5 \rfloor$
- **Verified equivalence**: 0.00% bpc difference between RNN and UM

The isomorphic UM uses only positive patterns, positive events, and positive integer support values. This is the starting point for interpretation.

2 The Tick-Tock Methodology

The core research methodology alternates between two phases:

2.1 Tock: Interpret the UM

1. Extract patterns from trained RNN weights
2. Find natural Event Spaces (coarser than doubled-E)
3. Name events and patterns semantically
4. Measure coverage: what fraction of performance is explained?

2.2 Tick: Train with Injected Knowledge

1. Inject discovered patterns into RNN weights
2. Train on data
3. Convert back to isomorphic UM
4. Return to Tock phase

Key insight: Interpretability and compression efficiency are the same problem. Understanding gained in Tock improves performance in the next Tick.

3 Validated Discoveries

3.1 ES1: Word Boundary Detector (h2)

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

Input patterns:

- Space $\rightarrow$ h2$^-$ with strength 8
- All letters $\rightarrow$ h2$^+$ with strength 1
- The 8:1 ratio creates binary switch via softmax

Performance:

- $\Delta h_2 > 0.95$ marks word-start with 99.6% accuracy
- At word-start: $\Delta h_2 = +1.87 \pm 0.11$
- At mid-word: $\Delta h_2 = +0.03 \pm 0.19$

Interpretation:

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

3.2 ES2: Syllable Momentum (h35)

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

Mechanism:

- Vowels send weak support to h_{35}^+
- Space sends support 2 to h_{35}^-
- Critical recurrent: $h_2^+ \rightarrow h_{35}^+$ with weight 0.612
- Self-connection: $W_{hh}[35, 35] = -0.184$ (decay)

Behavior:

- h_{35}^+ high early in word (syllable momentum)

- Decays over time due to negative self-connection
- Short words end with high h_{35}^{+}; long words with high h_{35}^{-}

Key finding: h_{35} encodes CV (consonant-vowel) syllable structure, not word identity. Words with same CV pattern have nearly identical h_{35} trajectories:

- CCV words (the, she): avg distance = 0.099
- CVC words (for, was, his, can): avg distance = 0.106
- VCV words (are, one, use): avg distance = 0.097

3.3 Character Class Event Spaces

Five character classes form natural Event Spaces with high within-class similarity:

Class	Count	Similarity
Digits (0-9)	10	0.9996
Punctuation (.,!?)	6	0.9992
Vowels (aeiou)	5	0.9884
Whitespace (space, tab, newline)	4	—
Other	232	—

Note: Consonants do NOT form a single ES—too diverse, likely split by phonetic properties.

3.4 Pattern Injection via SVD

SVD factorization of bigram patterns successfully initializes RNN weights:

Initialization	Initial bpc	After 10 epochs
Random	5.46	4.81
Pattern injection	4.47	4.53

Key result: 1 bit/char head start (18% improvement) without any training.
SVD component interpretation:

- Component 0 (97.5%): Frequency baseline
- Component 1 (1.1%): ASCII vs UTF-8
- Component 2 (0.3%): Letters vs Digits/XML
- Component 3 (0.2%): Bracket structure
- Higher components: UTF-8 internals, phonotactics

English bigrams have effective dimension ∼64 (top 64 components capture 99.9% of variance).

3.5 Neuron Clustering: 18 Natural Event Spaces

Correlation analysis reveals massive redundancy in 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

Implication: The 128 binary ESs (256 events) can collapse to ∼18 natural ESs.

Saturated neurons (carry no dynamic information):

- Always ON (7): h0, h10, h19, h34, h48, h103, h109
- Always OFF (6): h6, h23, h25, h32, h90, h91

4 Refuted Hypotheses and Corrections

4.1 The 53% Illusion

Initial claim: ES features explain 53% of compression (7.31 → 3.44 bpc).

Reality:

- Baseline had exploding weights from instability
- Only compared 0.1% of data (1M chars)
- Correct improvement: ≤3.7% theoretical maximum
- ES captures only 15.9% of byte-level Markov MI

Lesson: Validate against proper baselines; compute information-theoretic bounds first.

4.2 Spectral Radius Hypothesis (P2)

Prediction: W_{hh} eigenvalues should cluster near unit circle ($0.9 < |\lambda_{\max}| < 1.1$).

Result: $|\lambda_{\max}| = 2.52$ (dramatically outside unit circle).

Explanation: RNN does NOT use eigenvalue tuning for stability. Instead:

- W_{hh} is expansive (would explode without nonlinearity)
- Tanh activation clamps output to $[-1, 1]$, providing stability
- Memory mechanism is saturation, not eigenvalue decay

4.3 Memory Depth Prediction

Prediction: Memory should decay exponentially with $d_{\max} \approx 24/H_{\text{avg}} \approx 12$ characters.

Observation: Dependency roughly flat to 30 characters.

Possible explanations:

- Precision limit applies to gradients (training), not inference
- Tanh normalization prevents precision loss
- Information encoded redundantly across dimensions
- Trained W_{hh} learned efficient encoding

4.4 Word Identity Encoding

Hypothesis: Hidden states encode word identity.

Result: Word recognition from hidden states achieves only 4.9-6.4% accuracy (near random).

Explanation: A character-level model doesn't need word identity. It encodes $P(\text{next\_char}|\text{context})$, not "this is word X". Words are emergent, not explicitly represented. The "lexicon" is patterns of character transitions that covary.

5 Theoretical Framework

5.1 Unification: Q = λ

A central insight unifies four perspectives through the quotient-equals-luck principle:

Domain	Formulation	Meaning
Bayesian	$\lambda = 1/p$, $\Lambda = -\log p$	luck (inverse probability)
Thermodynamics	microstates shrink by λ	precision loss
Arithmetic Coding	interval shrinks by $p = 1/\lambda$	symbolic encoding
RNN	h encodes context, outputs p	prediction accumulation

The quotient $Q = |\text{prior}|/|\text{posterior}| = \lambda = 1/p$ unifies all four.

5.2 Time-Energy-Bits Relationship

From fundamental physics:

- Planck's relation: $E = hf$
- Landauer's principle: bits $\propto$ energy
- Energy-time uncertainty: bits $\propto h/\text{time}$

Depth limit interpretation: $d_{\max} = 24/H_{\text{avg}}$ simultaneously represents:

- A bit budget (information theory)
- An energy budget (thermodynamics)
- A temporal horizon (time)
- A precision limit (float32 has 24 mantissa bits)

5.3 AC-RNN Correspondence

Arithmetic coding and RNN hidden states are analogous:

- AC state: interval $[\text{low}, \text{high})$ shrinks as symbols encoded
- RNN state: $h \in \mathbb{R}^{128}$ evolves as text processed
- Both accumulate context over time
- Both hit precision limits (AC ∼32-64 bits; RNN ∼24 bits × 128 dims)

Key difference: RNN achieves stability through tanh saturation, not precision-limited arithmetic. The mechanisms differ even if the information-theoretic constraints match.

5.4 Pattern Depth is Free

Adding deterministic Event Space membership to inputs doubles effective pattern depth:

- Standard input: x_t (256 dims for byte)
- Augmented input: $(x_t, \text{ES}(x_t))$ (260 dims)
- ES membership is a lookup table (vowel, digit, punctuation)
- No learning needed—compile to lookup, inject into input

5.5 Factor Maps as Patterns on U^2

A theoretical contribution connecting:

- Factor maps $\pi : U \to U'$ are patterns on $U \times U'$
- Factor maps = embeddings = patterns (all live in U^2)
- Composition via tropical matrix multiplication
- Factor lattice structures model hierarchies

6 Current Coverage

6.1 Pattern Inventory

Pattern Type	Count
Input → Hidden (strength ≥ 1)	49
Hidden → Output (strength ≥ 1)	253
Hidden → Hidden (strength ≥ 1)	4,183
Total significant (excluding recurrent)	302

6.2 Event Coverage

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

6.3 BPC Attribution

Conservative estimate: Interpreted patterns explain <5% of the 5.69 bpc performance.

The bulk of compression comes from uninterpreted character-level statistics in the hidden→output weights. The model is primarily a character-level statistical predictor with word-boundary structure emerging from the space pattern.

7 Open Questions

7.1 From Previous Archives

1. **Why doesn't memory depth show predicted decay?** Observed flat to 30 chars, predicted 12.
2. **How do LSTMs select what entropy to carry?** Forget gate should correlate with local entropy.
3. **Can we inject longer patterns within precision budget?** Trigrams, 4-grams?
4. **What is the learning function ω in gradient terms?**

7.2 From Current Analysis

1. **How to collapse to natural ESs?** The 18 clusters need semantic names.
2. **What do h62, h3, h72 encode?** They receive strong space patterns but role unclear.
3. **How to represent words as character covariation?** Not lookup, but pattern structure.
4. **What patterns to inject for next tick?** Word endings, common bigrams?

8 Testable Predictions

Eight predictions from archive 20260131_5, with current status:

P#	Prediction	Status
P1	Random RNN memory decays exponentially	Untested
P2	W_{hh} spectral radius ≈ 1	**Refuted** ($\lvert\lambda_{\max}\rvert = 2.52$)
P3	LSTM forget gate $\sim$ entropy correlation	Untested
P4	SVD rank curve monotonic to $\sim$64	Validated (effective dim $\sim$64)
P5	Injection advantage shrinks with training	Validated (0.99 $\to$ 0.28 bpc)
P6	Hidden size = effective rank	Untested
P7	English bigram injection hurts non-English	Untested
P8	float64 doubles memory depth	Untested

9 Recommendations for Continued Analysis

9.1 Immediate Next Steps

1. **Name the 18 natural ESs**: Use correlation clusters to define coarser ESs with semantic labels.
2. **Analyze h62, h3, h72**: These receive strong space patterns; understand their role.
3. **Map hidden→hidden structure**: 4,183 recurrent patterns largely unexplored.
4. **Inject word-ending patterns**: Test if adding '-ed', '-ing', '-tion' patterns helps.

9.2 Medium-Term Goals

1. **Reach lexicon milestone**: Understand how $\sim$100 common words are encoded.
2. **Develop lift procedure**: Isomorphic UM $\to$ interpretable UM with fewer ESs.
3. **Build injection procedure**: Write new patterns without losing existing signal.
4. **Improve bpc**: Current 5.69 is far from 1.1 SOTA; each tick should improve.

9.3 Theoretical Work

1. **Formalize the tanh stability mechanism**: Why does saturation preserve information?
2. **Quantify precision loss in recurrent weights**: When does W_{hh} injection become possible?
3. **Develop hierarchical ES theory**: How do fine ESs (bytes) relate to coarse ESs (words)?

10 Conclusion

Eight days of research established:

1. **A working methodology**: The tick-tock cycle between RNN training and UM interpretation.
2. **An exact isomorphism**: The doubled-E UM matches RNN performance exactly.
3. **Two interpretable ESs**: Word boundary (h2) and syllable momentum (h35).

4. **Successful pattern injection**: SVD gives 1 bit/char head start.

5. **Massive redundancy**: 128 neurons collapse to $\sim$18 natural ESs.

6. **A unified theory**: Q = λ connects Bayes, thermo, AC, and neural nets.

7. **Key corrections**: The 53% claim was wrong; spectral radius hypothesis refuted.

The model is primarily a character-level statistical predictor. Word boundaries emerge from the space pattern; syllable structure from CV alternation. Words are not explicitly encoded—they are emergent patterns of character covariation.

Coverage remains low: $\sim$31% of events touched, $<$5% of bpc explained. The next phase should focus on understanding the hidden$\rightarrow$hidden recurrent structure and developing procedures for injecting lexical knowledge.