First Things First

Isolating the Order-Sensitivity Inductive Bias of LSTMs

Maurice Fifis — EEMCS, TU Delft — Supervisors: D.M.J. Tax, C.C.J. van Engelenburg — Committee: H.J. Griffioen

Background

- *Which architecture is best suited for a given problem?*
- A **problem archetype**: a dataset that isolates *why* one architecture wins, relative to all others [1]
- Synthetic benchmarks can reveal architectural differences obscured by complex real-world data [2]
- No problem archetype has been established for recurrent networks

Research Question

What characteristics of a synthetic dataset cause an LSTM to achieve lower classification error than non-recurrent architectures, and why?

LSTMs

- Process input **one timestep at a time**
- Maintain a **hidden state / cell state** — a memory of everything seen so far
- Gating controls what is written, retained, and read [3]
- Inductive bias: output depends on the **ordered sequence** of inputs

Method: Shared Evaluation Framework

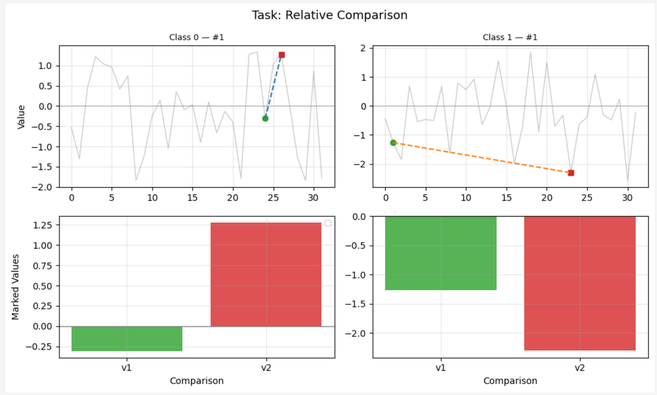
1. **Generate** — one synthetic dataset, shared 64-dim format
2. **Translate** — each of 5 architectures (LSTM, CNN, GCN, Transformer, MLP) reshapes the same input via a deterministic, parameter-free layer
3. **Train** — identical protocol: AdamW, BCE loss, early stopping
4. **Evaluate** — compare test accuracy, multiple seeds

Generate (shared 64-dim) → Translate (per-architecture) → Train (5 architectures) → Evaluate (test acc.)

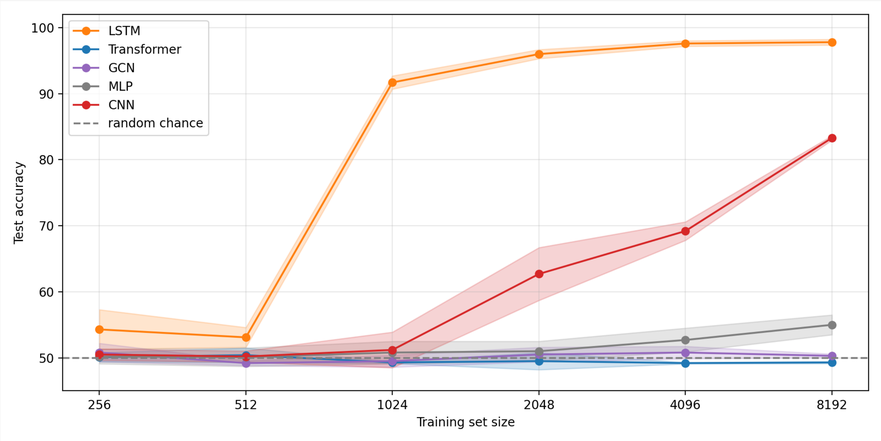
The Archetype: Relative Comparison

Task: sequence of 32 random values, two positions marked. *Is the value at the first marker larger than the second?*

The key argument: swapping the two marked values flips the label, but leaves the input statistically unchanged.

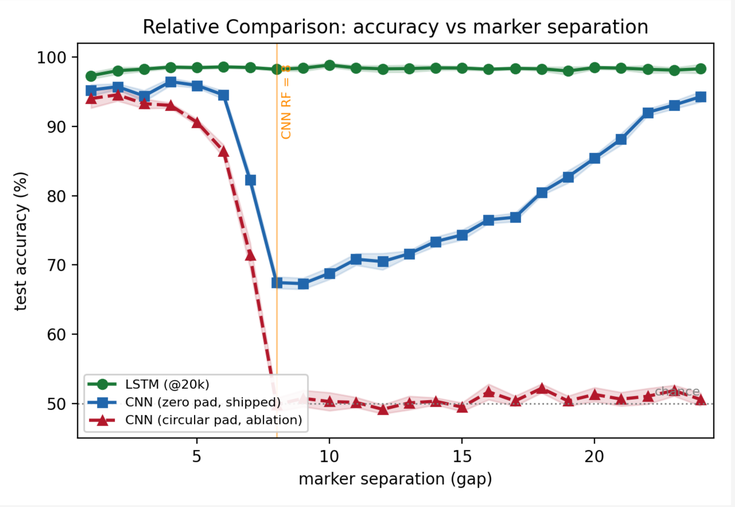


Result 1: Is the Gap Structural?



- LSTM converges cleanly to **97%** as training data grows
- Transformer and GCN stay at chance regardless of dataset size: their failure is structural, not a data shortage
- CNN improves with data (reaching ≈83% at N=8,192), but this is a separate story (Result 2)

Result 2: Why Does the CNN Partially Succeed?



- CNN compares correctly only **within its receptive field** (8 positions)
- Beyond that, its apparent recovery is a **padding artifact**: confirmed by a controlled ablation (circular vs. zero padding)

Conclusion

- Relative Comparison isolates **order-sensitivity**, not recurrence specifically
- Distills into 3 necessary design conditions: **no aggregate shortcut**, **order-dependence**, **practical learnability** — each validated via a dedicated failed-task ablation

Limitations & Future Work

- Distinguishing LSTM from other order-aware architectures needs a genuinely sequential archetype
- Deeper sequential task candidates were not learnable under the shared protocol
- Generalization to other recurrent variants (GRU, vanilla RNN) untested

References

[1] Duin, Tax, Jain — Classifier Problem Archetypes
[2] Greydanus, Kobak — Scaling Down Deep Learning with MNIST-1D
[3] Hochreiter, Schmidhuber — Long Short-Term Memory, 1997

Contact: m.v.d.f.fifis@student.tudelft.nl