Neuro-Symbolic Systems and the Challenge of Artificial General Intelligence

Florin Leon

"Gheorghe Asachi" Technical University of Iași
Romania

Frontiers in Artificial Intelligence: Challenges and Perspectives
Smart Diaspora 2025, Cluj-Napoca, Romania

Outline

- Motivation: can reasoning LLMs be improved?
- Variable binding and indirection
- Compositional generalization
- Neuro-symbolic and brain-inspired architectures
- Neural cell assemblies
- Discussion, conclusions and open questions

Reasoning LLMs

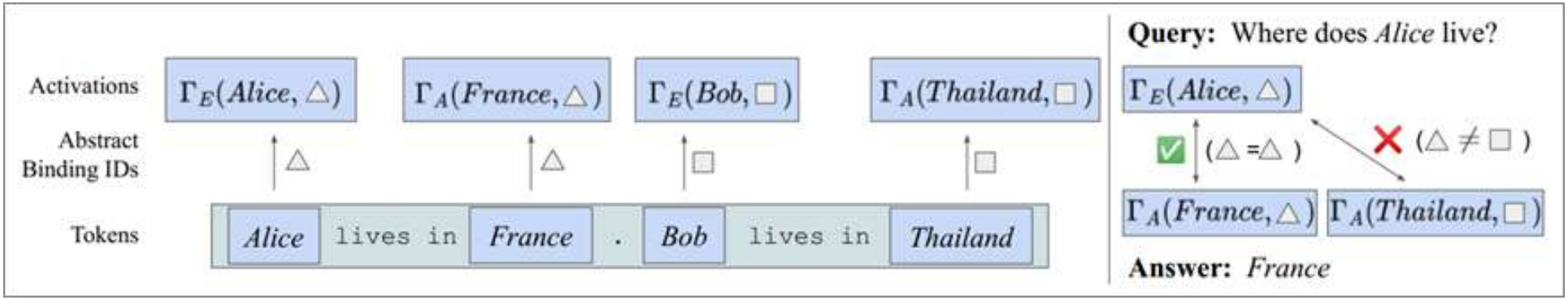
- Reasoning Large Language Models (or Large Reasoning Models) can now solve multi-step academic problems by simulating human-like workflows
- Methods like chain of thought make models more powerful and their external behavior more transparent
- They remain resource intensive and their internal representations are difficult to interpret
- Brain-inspired principles could improve scalability and reasoning
- We focus on mechanisms that could improve reasoning in neural systems and LLMs
  - Less focus on logic, which may be a learned, constrained mode of thought

Variable Binding

- Variable binding is the problem of linking roles or properties to the correct entity, without confusing or blending multiple items together
- Role-filler examples:
  - Maria moves her hand (subject/agent → Maria, object/patient → hand)
  - A person has an age; Alex is 30 (person → Alex, age → 30)
- Similar to the substitution of variables with symbols in first-order logic
- Feature binding:
  - A blue circle and a red rectangle vs. A red circle and a blue rectangle

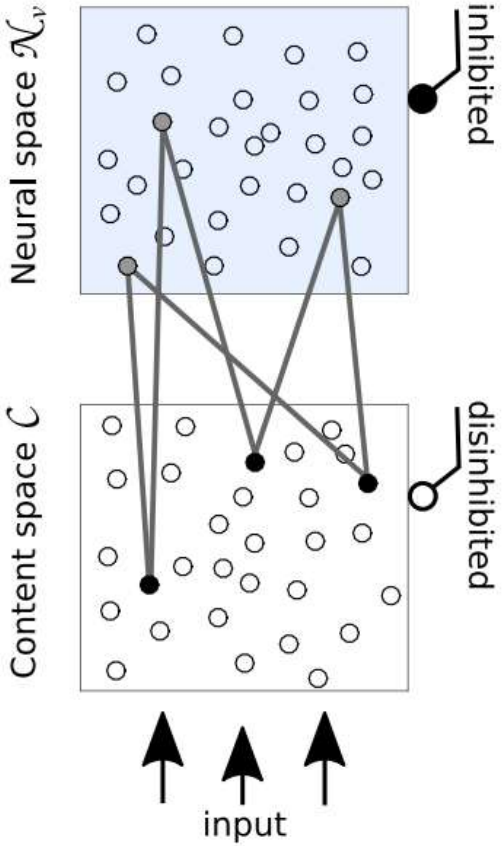
Binding ID Vectors

- Larger language models develop binding ID vectors between entities and attributes
- These vectors enable correct associations within contexts, support generalization between tasks, and form a continuous subspace



Pointer-Style Mechanisms

- Assembly pointers can bind variables
  - A role assembly activates a target assembly through fast Hebbian plasticity
  - This supports indirection and memory indexing using, e.g., SNNs
- In the brain, pointer-style mechanisms may be supported by the prefrontal cortex and basal ganglia through indirection
  - Role neurons encode the locations of filler neurons



Kriete et al., *Indirection and Symbol-Like Processing in the Prefrontal Cortex and Basal Ganglia*, Proceedings of the National Academy of Sciences, 2013

Legenstein et al., *Assembly Pointers for Variable Binding in Networks of Spiking Neurons*, Workshop on Cognitive Computation – NIPS 2016

Compositionality

- Compositionality is the problem of forming complex representations whose meaning arises from the meaning of simpler parts and their combination
- Example: SCAN dataset
  - Training: run → RUN; run twice → RUN RUN; jump → JUMP
  - Testing: jump twice → JUMP JUMP
- Neural networks typically require extensive training data; humans generalize with less data
- Many deep networks have strong performance, but may fail tasks that require rule-based composition of known elements into new combinations

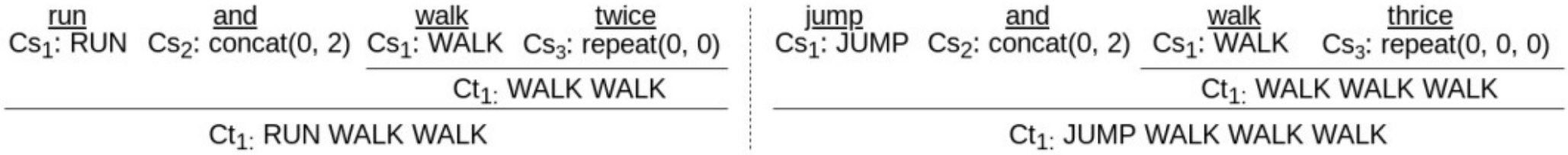
Lake & Baroni, *Generalization Without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks*, ICML 2018, https://github.com/brendenlake/SCAN

Meta-Learning for Compositionality

- MLC trains a standard Transformer to generalize better
- Each episode presents a new artificial mini-language
  - Training: *dax* → ●; *dax fep* → ● ● ●
  - Testing: *lug fep* → ● ● ●
- Grammar and rules vary per episode to encourage generalization not memorization
- MLC matches human compositional performance and replicates typical human error patterns

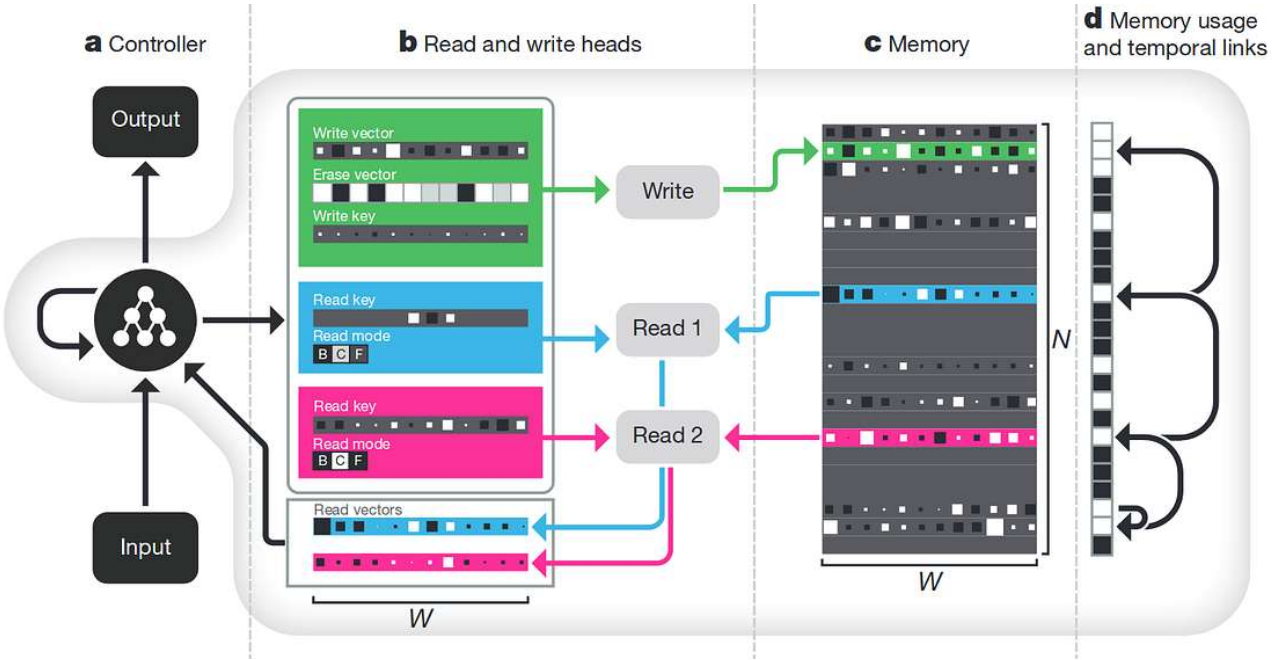
Neural Symbolic Stack Machine

- A neural controller (LSTM with attention)
  - Predicts the sequence of operations to generate the output
  - Learns when to apply each operation, rather than producing tokens directly
- A stack-based interpreter
  - Executes symbolic operations like *push*, *pop*, *concat*
  - Builds the output recursively
- NeSS is a fully differentiable model



Differentiable Neural Computer

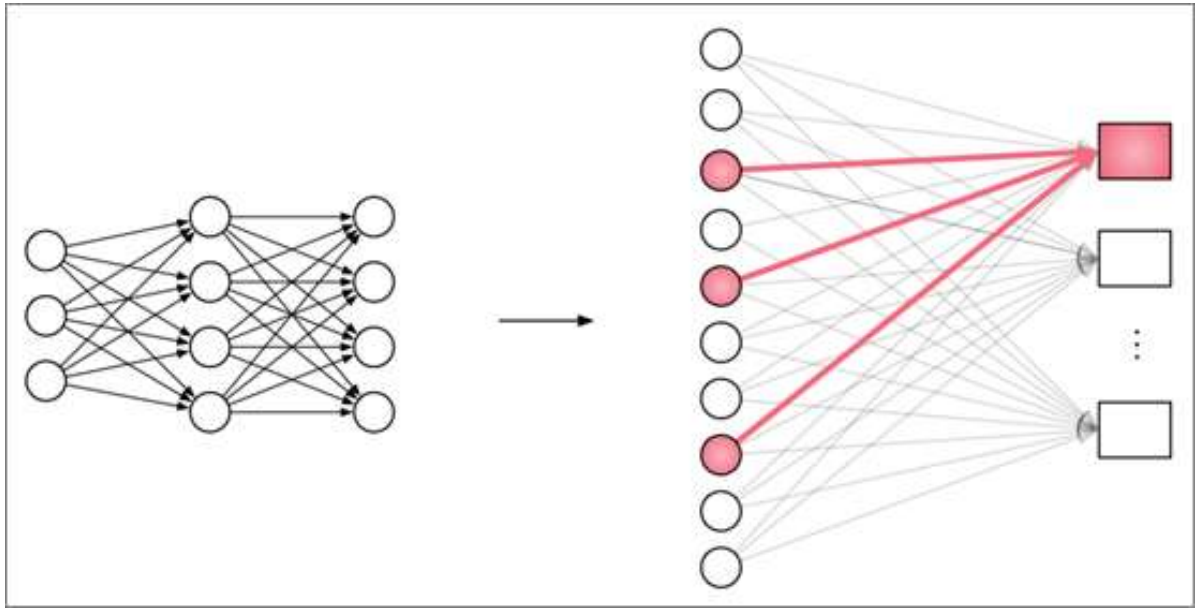
- A recurrent neural network augmented with an external differentiable memory, trained end-to-end to learn algorithmic behavior
- Memory with soft attention: content-based addressing
- Can solve tasks like pathfinding in graphs, array sorting
- Generalizes to variable-length inputs and unseen sequences



Graves et al., *Hybrid Computing Using a Neural Network with Dynamic External Memory*, Nature, 2016

FlyVec Embeddings

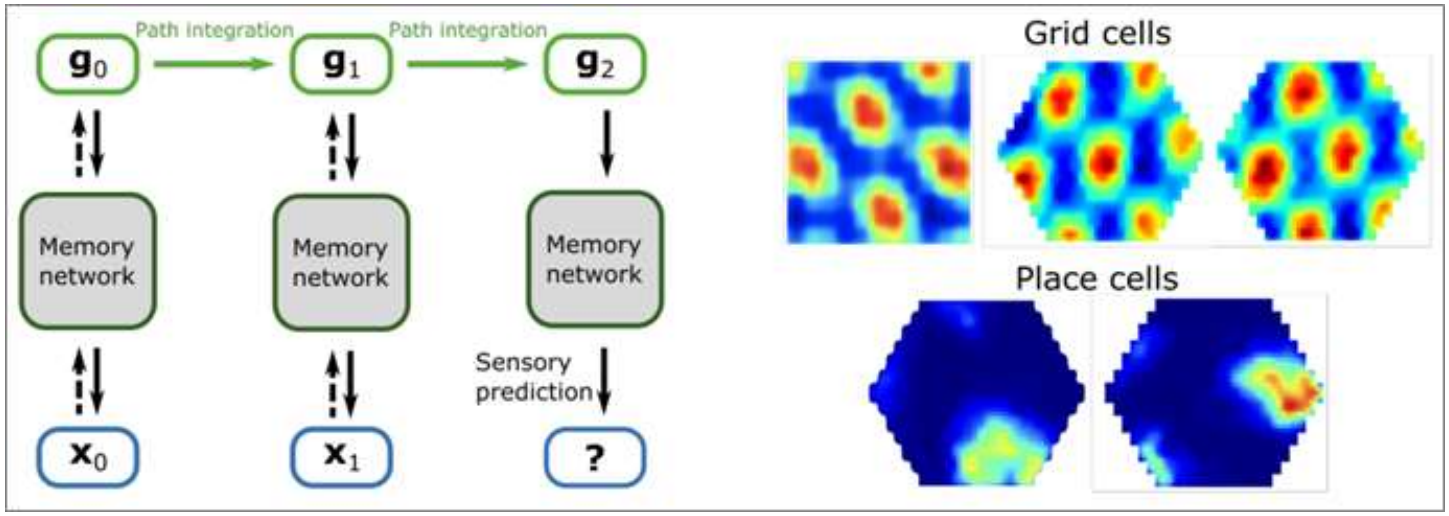
- Embedding is the problem of mapping discrete symbols to continuous vectors that capture meaning and potential relationships
- Sparse binary embeddings for words and sentences can be created using inspiration from the fruit fly olfactory system with a *k*-winners-take-all mechanism



Liang et al., *Can a Fruit Fly Learn Word Embeddings?* ICLR 2021

Transformers and TEM

- The Tolman-Eichenbaum Machine (TEM) separates *where* (locations, transitions) and *what* (sensory observations) and binds them conjunctively (hippocampus)
- TEM-t is a Transformer with learned recurrent positional encodings, functionally equivalent to TEM

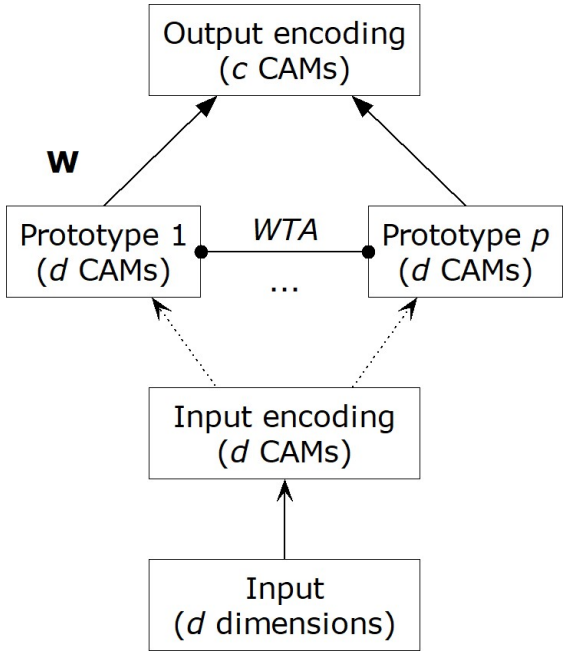


Whittington et al., *The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory Through Generalization in the Hippocampal Formation*, Cell, 2020

Whittington, Warren & Behrens, *Relating Transformers to Models and Neural Representations of the Hippocampal Formation*, ICLR 2022

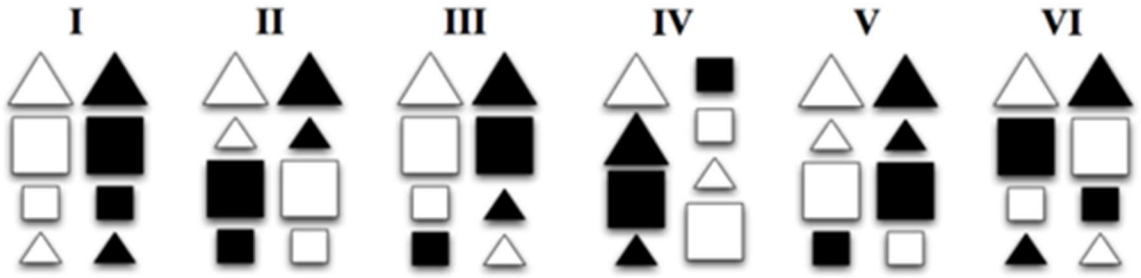
Classification with Cell Assemblies

- A neuro-symbolic classification algorithm inspired by the SUSTAIN algorithm in cognitive psychology
- Internal representation based on neural cell assemblies, an intermediate level between neurons and symbolic concepts
- Prototypes for similarity-based classification, rule extraction for explicit reasoning



CAM = cell assembly module
WTA = winner takes all

The Six Shepard Problems

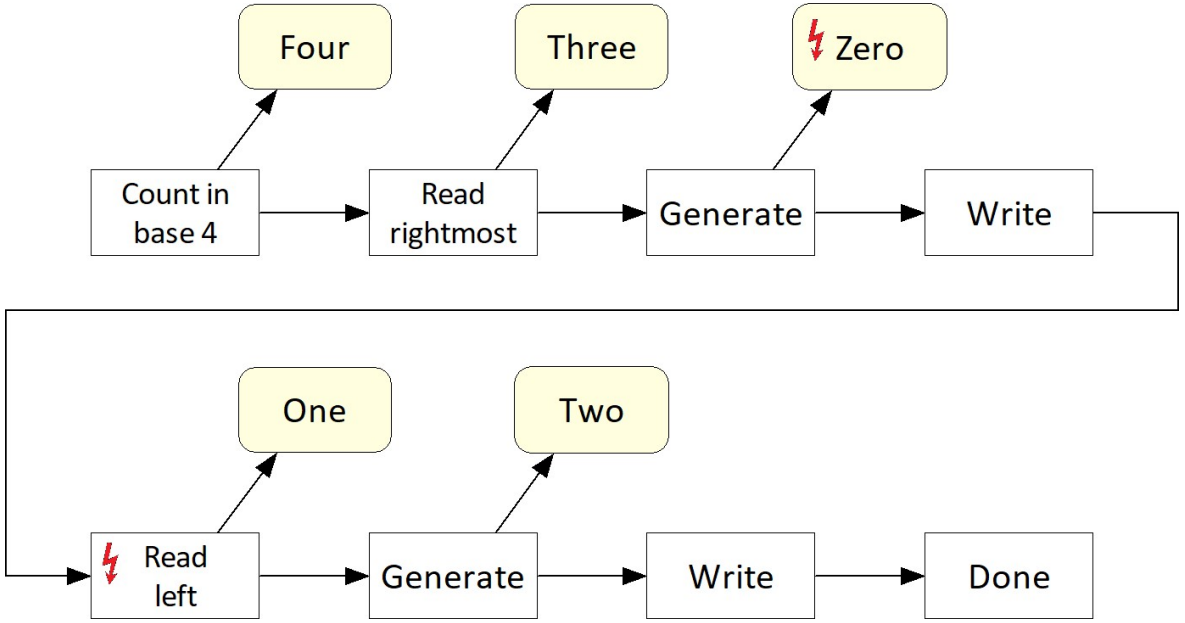


color
shape
size

I	0.00 (0.95)	0.41 (0.03)	0.55 (0.02)	C0
	1.00 (0.97)	0.51 (0.00)	0.57 (0.03)	C1
II	1.00 (0.49)	1.00 (0.49)	0.45 (0.01)	C0
	0.00 (0.50)	0.00 (0.50)	0.52 (0.00)	C0
	1.00 (0.50)	0.00 (0.50)	0.47 (0.01)	C1
	0.00 (0.50)	1.00 (0.50)	0.52 (0.00)	C1

Learning to Plan with Cell Assemblies

- Learning to count, like a child's learning process, by imitation learning
- First learning the concepts of digits and successors in base 10
- Then learning to count in a different base (e.g., 4) in general, possibly to infinity, not within a specified range
- Learning based on surprises, e.g., for base-specific carry-over operations



Discussion

- Explicit role representations in the embedding space could let the model assign stable placeholders to entities for reasoning tasks
  - These placeholders could support variable-like substitution through fast plasticity
- Pointer-like indirection could separate roles from values, and this could enable flexible recombination and better generalization
  - This could also encourage creativity: familiar elements could be reused in new contexts
- The separation between function and content could improve interpretability, enable stable editing, and support complex or nested tasks more reliably
  - E.g., a neural controller with symbolic components like memory or stacks could support algorithmic, recursive, and hierarchical processing

Conclusions

- Biological systems achieve efficient, compositional reasoning with strong generalization, in contrast to the high cost and low interpretability of current LLMs
- Incorporating brain-inspired principles may lead to neural architectures that:
  - Reason more transparently
  - Generalize more reliably
  - Approach the adaptability of human cognition

Some Open Questions

- Are we at the beginning of a socio-economic and industrial revolution driven by AI?
  - Probably, yes
  - Will human intelligence become obsolete? Will it matter?
  - Will society split between people who understand science and those who do not?
- What will be the impact of Artificial Superintelligence (ASI) and Artificial General Intelligence (AGI)?
  - Will people thrive and focus on the most meaningful matters?
  - Will people become lazy and disconnected and lead society to collapse (mouse utopia experiments)?