

17/01/2020

From Brain Signal to Language [Paul Smolensky]

Paul Smolensky

What we are at:

→ New kind of existence to co-exist with

→ "Final Revision" story to Claude

↳ NN, "neurons connected with each other" "Connectionism"

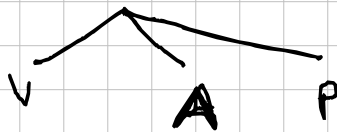
SYMBOLS: INPUT & OUTPUT → symbols, where
↳ recurrent activation vector

↳ How does it look from the p.o.v of COGNITIVE science?

COGNITION $\approx$ COMPUTATION

A Pendulum...

Discrete symbolic structures (1955)
(manipulated by rewrite rules)



Numerical Vectors (1985)

PDP generation

Discrete symbol structures 1985
A = assigned probabilities

Numerical Vectors

Deep Learning generation (2005)

What kind of computational system is cognition?

↓
symbolic computation to understand models

↘ Neural Comp, LLM

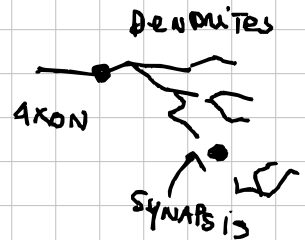
What is The secret of human intelligence?

PROGRAM: MIND

- REASONING
- MATH
- NLP

HARDWARE: BRAIN

- General properties
- Detailed properties



Discovering The implicit symbolic structure of LLM

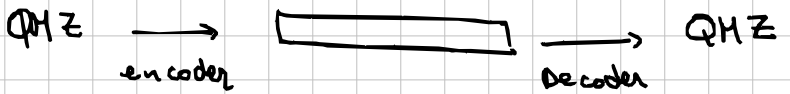
(Tom McCoy)

Dog chase cats
↑
Cats chase Dog

1 0 1 0 0 1

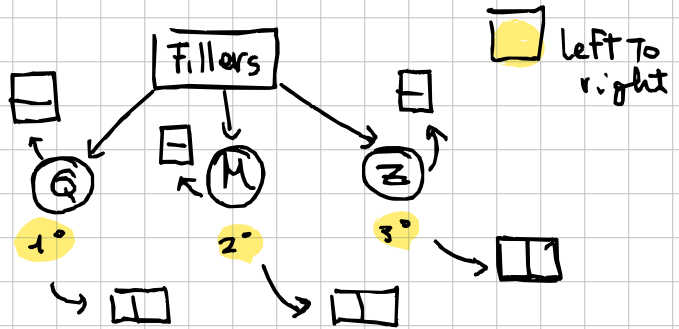
1 0 1 0 0 1

HP: They encode Tensor Product Representation



OR... ZMQ

Single
Networks
Bottleneck



Tensor products of $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\rightarrow$ LINEAR MAP

NN are implicit TPRrepresenters

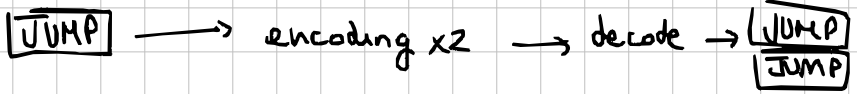
We Feed The TPRs To a decoder To check

MLP
RNN
Transformer

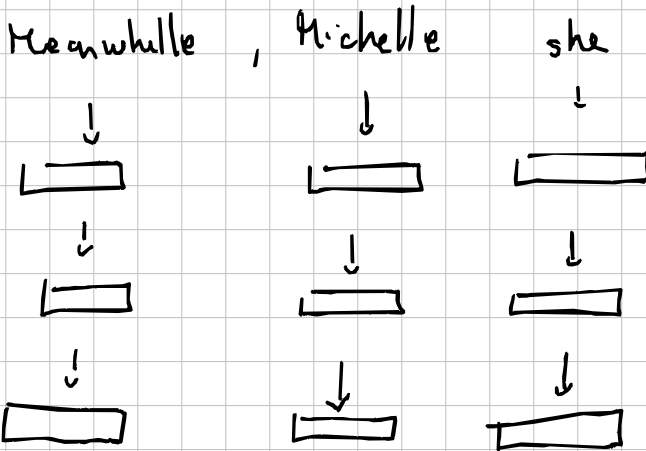
} Turns out BIDIRECTIONAL WORK 100% $\forall$ architecture

UNcovering TPR in Embed Representation, (Tom McCoy)

Jump twice

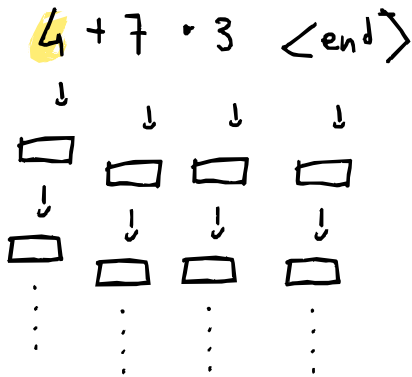


Prompting Claude



D. TPR appear in LLM as well?

- Arithmetic
- Logic
- Linguistic Transformations
- ...



Sequential
SYNTACTIC

We construct
hypothesis on
STRUCTURE

(Position, Pos ...)

25

TPR in GA-OSS

Summe performances
on these tasks

Can we do causal intervention? ✓

The clever doctor helps the lawyer

- clever: subject-adjective + clever: object-adjective

TPR explains other interpretability methods ✓

Also TPR connect

NN theory
(BRAIN)

→

SYMBOLIC
RULE THEORY
(MIND)