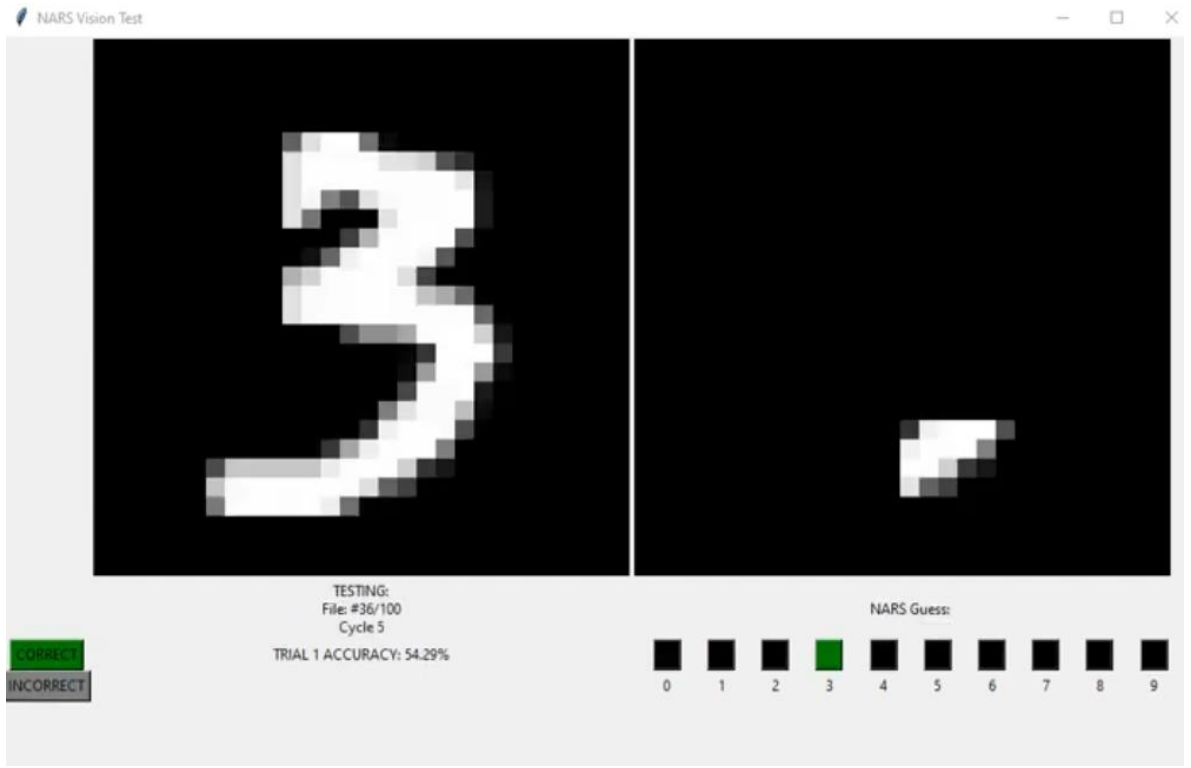
Visual Perception Experiment with NARS: MNIST Digit Recognition

By: Christian Hahm

Temple AGI Team

Temple University



Male Northern Cardinal (state bird of Ohio, USA)

Perception through Reasoning?

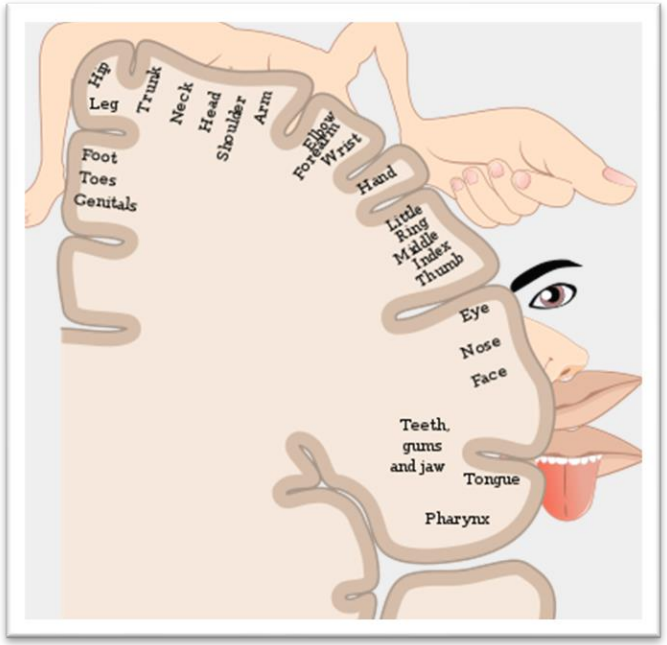
At the current stage, NARS vision requires a pre-processing module, *e.g.*, convolutional neural network, to detect high-level objects (*e.g., "cardinal"*) from low-level pixels (*the image on the left*).

Once NARS has the high-level "abstract" concepts, they can be used in high-level **logical reasoning:**
e.g., deduction:
{<cardinal→bird>., <bird→animal>.}
⊢ <cardinal→animal>.

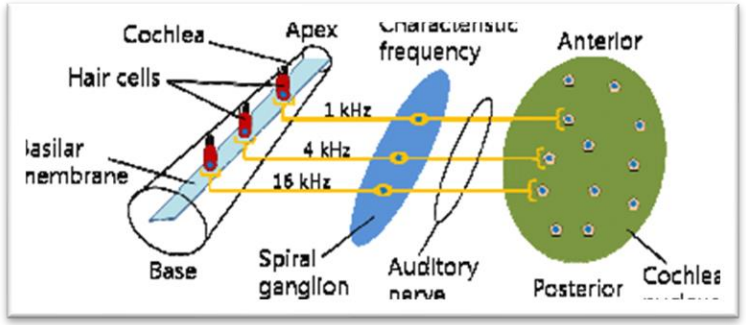
But what about raw sensor data (e.g., the pixels of an image)? Can the system do logical reasoning with sensations to build abstract concepts on its own? After all, NARS can natively compute **abstractions** and **generalizations** of high-level concepts, why not with sensor data?

Research Question: Can NARS make **actionable predictions** by natively composing ***high-level concepts (e.g., compounds)*** from ***low-level sensations***, without the need for a separate vision module?

Background



Somatotopic (Touch) Map



Tonotopic (Hearing) Map

Human Sensation & Perception

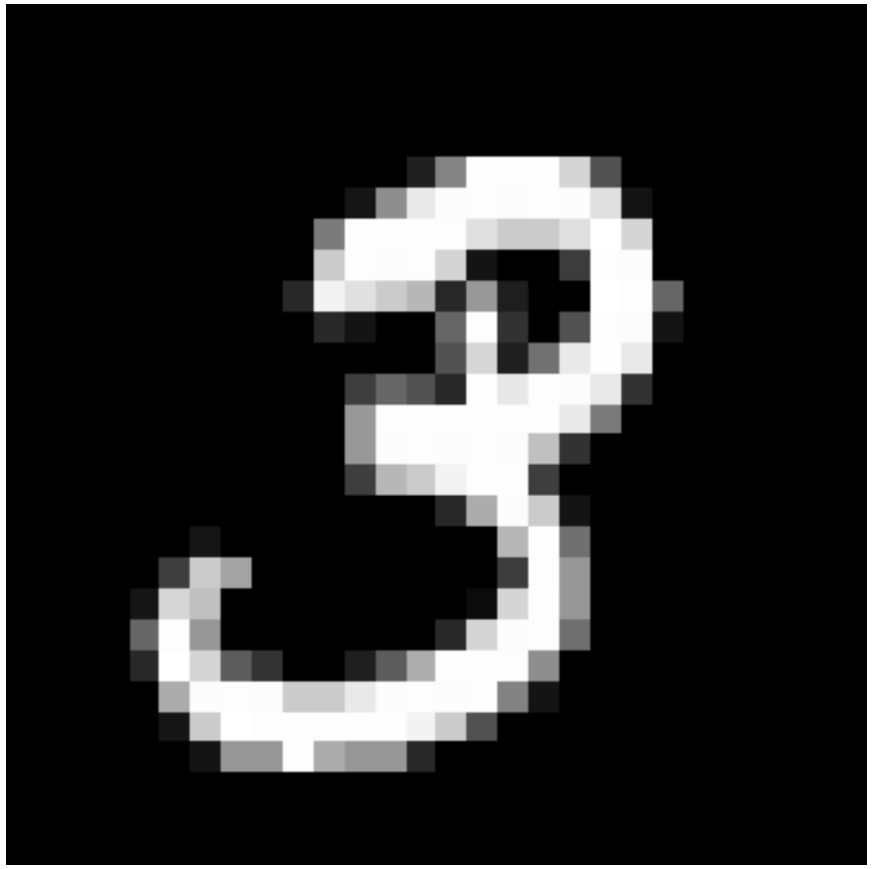
- Humans have more than 5 sensory modalities (e.g., vision, touch, hearing, etc.). For each modality, there are neurons which detect **specific physical signals from the world** (*e.g.,* red photoreceptor, mechanical pressure receptor, high pitch, etc.).
- Sensory neurons can be activated strongly, weakly, or not at all, depending on the stimulus pattern. These "sensations" are the nervous system's "inputs".
- Each neuron is located at a unique location in space. Sensory neurons connect to the brain in a way that **preserves their relative spatial layout,** also called a **topographic mapping** (see [*Wolfe et. al, 2006*], and images to left)
- The brain apparently uses topographic maps to understand the spatial layout of its sensations. **If NARS is analogous to the brain, can we modify NARS to accept sensory topographic maps, then test NARS performance?** That's what this experiment is about.

Info about the MNIST Digit Dataset

- MNIST digit recognition is a famous test for computer vision; system succeeds by correctly identifying the digit in the image (*for example, see right*)

- Dataset containing labeled images of handwritten digits (0-9)
  - 28 pixels x 28 pixels = 784 pixels

- MNIST is a decent test for NARS perception. We treat each image like a topographic map of 784 **“low-level”** sensory photoreceptor activations, from which NARS can infer the **“high-level”** object class (the digit’s label).



MNIST Image of digit 3

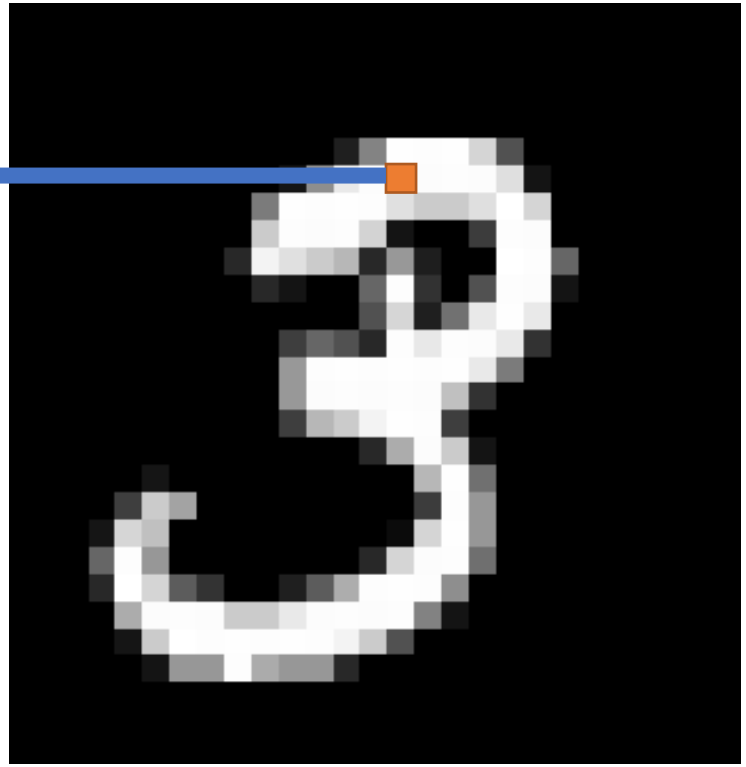
NARS
Modifications

Topographic Mapping in NARS using "Spatial Buffer"

- Treat each pixel as a vision sensor *instance (e.g, like a single photoreceptor).* The *instance* inherits a certain *property* (the property the sensor detects, *e.g., "bright", "red", "green", "vertical edge", etc*.).

- Here we use [BRIGHT] property to indicate the pixel's brightness / intensity.

- Encode each sensor's activation level as an atomic Narsese *Event*, where:
  - *frequency* = intensity of the sensation (0 = no activation, 1 = full activation)
  - *confidence* = unit amount of evidence = $\frac{1}{1+k}$, where *k* is the evidential horizon
  - ***e.g.,* <{Pixel_x14_y06} → [BRIGHT]>. :|: <0.95, 0.5>**



- **Change to NARS architecture:** store the events in a 2D "**Spatial Buffer"**, which retains the spatial layout of the input events, so they can be grouped by location. Also stores a reference to them in a **Bag** using the **event's truth expectation** as the **item's priority** so they can be randomly selected (i.e., items near *expectation=1.0* are more likely to be selected).

Spatial Buffer and “the Spotlight”

- First, NARS probabilistically selects an atomic event from the Bag; this is the “pivot” event

- Then, since spatially near events are more likely to be related than spatially distant events, NARS can **group events** using a **spatial window** around the pivot event

- In psychological literature, this is called the **Spotlight model of attention;** some argue the spotlight can vary in size according to certain factors, in the **Zoom-Lens Model of Attention** (see [Eriksen and St. James, 1986], [Cave and Bichot, 1999] for further reading)

- Multiple events selected in the window are combined into a higher-order statement using conjunction (AND):

E.g., 2 events:

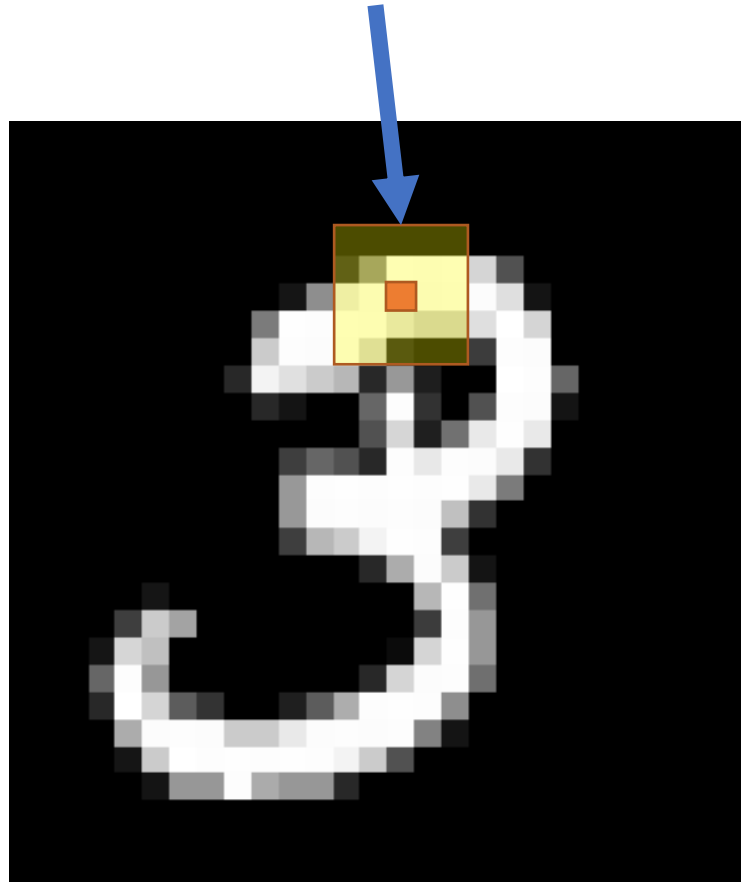
<({Pixel_x14_y06} → [BRIGHT]> ∧ <({Pixel_x15_y05} → [BRIGHT]>. :|:

This statement is equivalent to the first-order *intensional intersection* (*aka extensional union*):

= <({Pixel_x14_y06} ∪ {Pixel_x15_y05}) → [BRIGHT]>. :|:

= <({Pixel_x14_y06, Pixel_x15_y05}) → [BRIGHT]>. :|:

Spotlight of attention

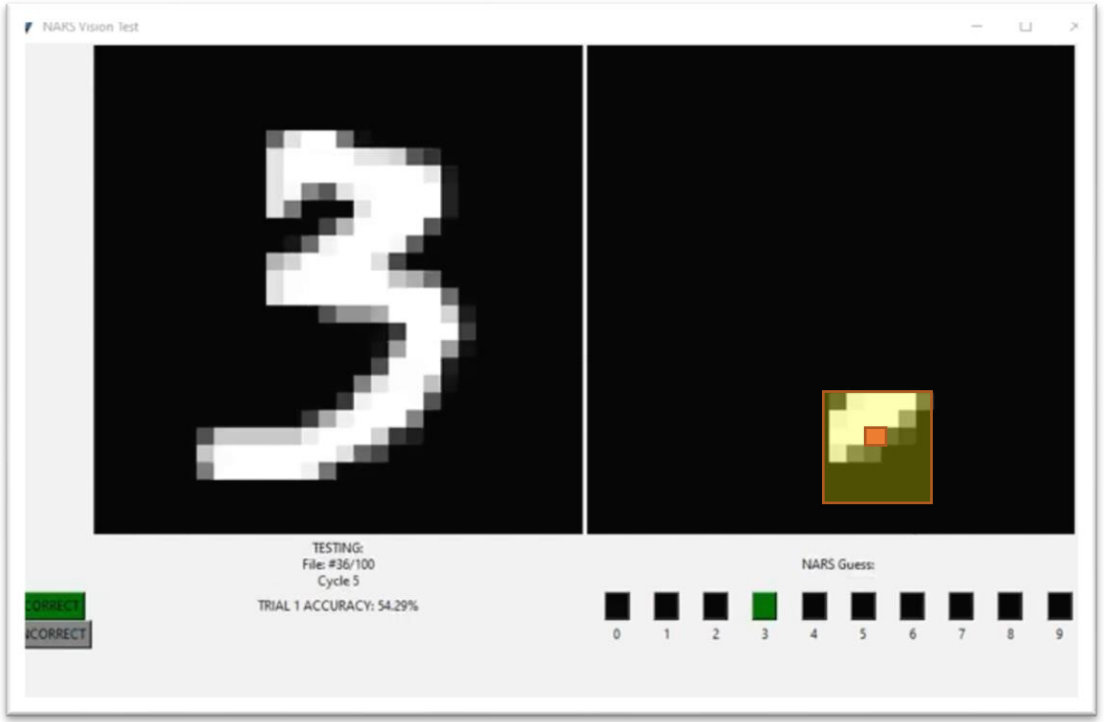


e.g., a 5x5 window, would compose up to 25 atomic events into a single compound event

Experiment Methodology

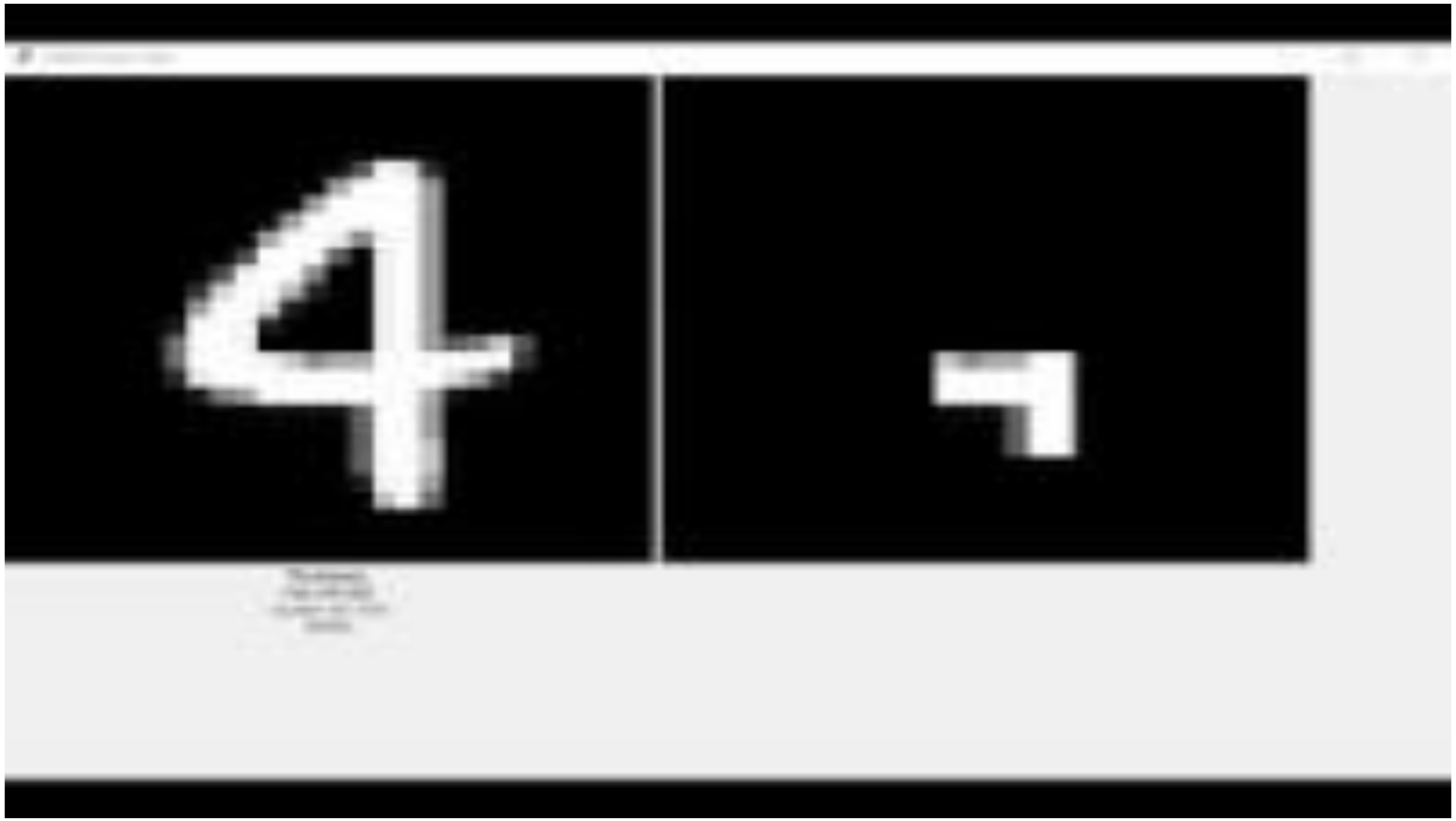
Experimental Test Method

- **IDEA:** NARS will accumulate evidence as it scans the image and composes visual features. NARS will continuously make predictions of image class based on the features it sees; these predictions accumulate in memory.
- **HYPOTHESIS:** When the predictions for a digit accumulate strongly, NARS derives executes an operation to guess that digit. By grouping multiple events into compounds, the compound is specific enough to predict the digit's class, but generic enough to recognize variations of the digit (compared to a single pixel which is too generic, or an entire image which is too specific).
- **SETUP:** Each experimental trial is divided into Training Phase & Testing Phase. First, the system is shown labeled images in the Training Phase. Then, its knowledge is tested in the Testing Phase
  - **In Training Phase:** NARS is simply let to observe the training image for some number ***T*** of cycles, and input with a "label event" informing NARS which digit it is seeing (e.g., **<Digit5 → [seen]>. :|:).**
    During this phase, NARS temporally associates visual features with the digit label event using **induction:**
    (e.g., **<<{Pixel_x14_y06} → [BRIGHT]> =/> <Digit5 → [seen]>.**)
  - **In Testing Phase:** NARS is given seed goals to "execute operation for digit X if digit X is seen" (e.g.,. **G5= <<Digit5 → [seen]>,^pressDigit5>!**) for all 10 digits.
    Therefore, when NARS predicts event **E = <Digit5 → [seen]>,** NARS can use **deduction** to derive evidence for operation goal:
    {G5, E} ⊢ ^pressDigit5!
- **SUCCESS CRITERIA:** NARS gets one chance to answer for each image. It answers by executing an **operation** (*e.g., ^pressDigit3*), and the answer is then recorded as "**correct**" or "**incorrect**".
  - **TIMEOUT:** If the system does not recognize the image features, it might not execute any operation at all, in which case NARS was marked "**incorrect**" after a **timeout period of 3000 working cycles**.



In this screenshot, NARS correctly executes a '3' operation, indicating that it recognizes the digit '3' in the image based on the image's features

Experiment video



Results & Analysis

Accuracy Results

Test Name	*Number of Train/Test Images*	*Training Cycles (per image)*	*Trial 1*	*Trial 2*	*Trial 3*	**Overall Avg. Accuracy**
Binary Memorization [0, 1]	10, 5 per digit	150	100%	100%	100%	**100%**
Digit Memorization [0 – 9]	10, 1 per digit	1500	100%	100%	100%	**100%**
Binary Classification [0, 1]	30/90	750	97.78%	98.89%	96.67%	**97.78%**
Digit Classification [0 – 9]	300/100	125	48.0%	43.0%	40.0%	**43.66%**

Personality Parameters

Test Name	*Evidential Horizon (k)*	*Cautiousness (T)*	*Priority Mask Focus*	*Event Time Decay*	*Desire Time Decay*
Binary Memorization [0, 1]	7	0.600	20.0	0.070	0.999
Digit Memorization [0 – 9]	22	0.582	0.159	0.560	0.999
Binary Classification [0, 1]	22	0.582	6.380	0.913	0.832
Digit Classification [0 – 9]	1	0.65	0.935	0.95	0.95

Important Results

- Results:
  - NARS performed memorization perfectly.
  - It could classify never-before-seen images of binary digits about 98% of the time

Test Name	*Number of Train/Test Images*	*Training Cycles (per image)*	*Trial 1*	*Trial 2*	*Trial 3*	**Overall Avg. Accuracy**
Binary Memorization [0, 1]	10, 5 per digit	150	100%	100%	100%	**100%**
Digit Memorization [0–9]	10, 1 per digit	1500	100%	100%	100%	**100%**
Binary Classification [0, 1]	30/90	750	97.78%	98.89%	96.67%	**97.78%**
Digit Classification [0 – 9]	300/100	125	48.0%	43.0%	40.0%	**43.66%**

Table 1. Accuracy Results recorded for the NARS MNIST Digit Memorization and Classification Tests (3 trials).

- **Digit Classification** - the hardest test, identify new images of digits 0 to 9. On 1 trial, out of a test set of **100 images**, NARS **correctly classified 48% of images:**
  - NARS classified almost half of all images correct. Much better than random chance (10%), so the method shows promise
  - NARS simply used composition, induction, and deduction rules

- **Personality parameters:**
  - **0.95 decay rate** for both events and goals
  - **Evidential horizon** was standard, ***k = 1***
  - **Decision-making threshold** was moderate, ***T = 0.65***

Test Name	*Evidential Horizon (k)*	*Cautiousness (T)*	*Priority Mask Focus*	*Event Time Decay*	*Desire Time Decay*
Binary Memorization [0, 1]	7	0.600	20.0	0.070	0.999
Digit Memorization [0–9]	22	0.582	0.159	0.560	0.999
Binary Classification [0, 1]	22	0.582	6.380	0.913	0.832
Digit Classification [0 – 9]	1	0.65	0.935	0.95	0.95

Analysis of Results

- The spatial grouping of visual features for compositional inference allowed NARS to correctly predict the class of novel images
- The system achieved higher accuracies when modified to:
  - Use **static-sized** attention windows (5x5), rather than variable **randomly-sized** attention window. In the future, attention window should be guided by top-down predictions
  - Given access to select **disjunctions of events** from a Spatial Buffer, as a form of "max-pooling" for robustness to small local translations:
    - **E.g,**

// Say we have a 4x4 feature map:

$$\begin{bmatrix} \neg E_{13,14} & E_{13,15} & \neg E_{13,16} & \neg E_{13,17} \\ \neg E_{14,14} & E_{14,15} & \neg E_{14,16} & \neg E_{14,17} \\ \neg E_{15,14} & E_{15,15} & \neg E_{15,16} & \neg E_{15,17} \\ \neg E_{16,14} & E_{16,15} & \neg E_{16,16} & \neg E_{16,17} \end{bmatrix}$$

// Nearby features can be combined disjunctively as a form of max pooling (in this case, 2x2 pooling wit
stride 2, resulting in a 2x2 pooled feature map from which elements can also be selected):

$$\begin{bmatrix} <E_{13,14} \vee E_{13,15} \vee E_{14,14} \vee E_{14,15}> & \neg<E_{13,16} \vee E_{13,17} \vee E_{14,16} \vee E_{14,17}> \\ <E_{15,14} \vee E_{15,15} \vee E_{16,14} \vee E_{16,15}> & \neg<E_{15,16} \vee E_{15,17} \vee E_{16,16} \vee E_{16,17}> \end{bmatrix}$$

Limitations
and
Future Improvements

Limitations and Future Directions

- The system could not choose its own attention window, rather it was randomly sized. Especially it would be useful if NARS could use predictions using the existing concepts in memory (*e.g., NARS see a red pixel, so NARS predicts it will see a cardinal and should look for beak features to confirm*)

- The only feature map available to NARS was an "intensity" map. Humans have neurons which detect specific edge orientations (*e.g., vertical, horizontal, 30^o,45^o ,etc.*) across the visual field, which could also be represented as an activation map, though which exactly which visual features to support must be investigated.

- **Related question:** How can NARS detect complex features from an RGB image? *e.g.*, human brains have "vertical edge detectors", can NARS detect vertical edges using Boolean operators?

- The explicit spatial window used here might be a naïve approach, vs. implicit and parallel spatial composition which could be achieved using *e.g.,* neural networks. Still, these results show the promise of spatial grouping and evidence-based reasoning for sensory understanding.

- **Further research directions:** Here NARS was trained and tested in a passive image classification task. Future tests could take place in an active context by testing NARS goal performance in a real-time interactive environment.

References

- [**Cave and Bichot, 1999**] *Visuospatial attention: Beyond a spotlight model*
- [**Eriksen and St. James, 1986**] *Visual attention within and around the field of focal attention: A zoom lens model*
- [**Wang et al., 2022**] *A Model of Unified Perception and Cognition*
- [**Wolfe et al., 2006**] *Sensation & Perception*

Image Sources

- Cardinal: https://en.wikipedia.org/wiki/Northern_cardinal#/media/File:Male_Northern_Cardinal_in_Hudson,_Ohio.jpg
- Homunculus: https://en.wikipedia.org/wiki/Cortical_homunculus
- Tonotopy: https://www.researchgate.net/publication/257775085_Sound_quality_analysis_of_a_passenger_car_based_on_electroencephalography/figures?lo=1

END