Visual concept formation using a dynamic system

Weibo Gong and Bradley Zylstra
University of Massachusetts
Amherst, MA 01002, USA

Abstract— Spatial signal processing algorithms often use pre-given coordinate systems to label pixel positions. These processing algorithms are thus burdened by an external reference grid, making the acquisition of relative, intrinsic features difficult. This is in contrast to animal vision and cognition: animals recognize features without an external coordinate system. We show that a visual signal processing algorithm with content-based addressing is not only important for animal vision, but also fundamental for concept formation. In this paper we start with a visual object deformation transfer experiment. We then formulate an algorithm that achieves deformation-invariance with relative coordinates. The paper concludes with the implications for general concept formation.

I. Introduction

Recent developments of Large Language Models heated up the historical debate in AI about how a machine should generate new concepts from data corpus. Furthermore, could the concept formation mechanism enable a thought process via recursion? As early as 1976 the recursive formation of abstract concepts in thought processes had been noted by the great scientist Stanislaw Ulam:"There must be a trick to the train of thought, a recursive formula. A group of neurons starts working automatically, sometimes without external impulse. It is a kind of iterative process with a growing pattern. It wanders about in the brain, and the way it happens must depend on the memory of similar patterns" [9].

To address this fundamental issue it might be useful to go back to the more specific mechanism of generating concepts from visual objects in the animal brains. When a visual object deforms, the animal brain acquires the concepts of the deformations such as rotation or stretching.

The animal vision system recognizes deforming objects effortlessly, whereas machine learning algorithms have yet to match up. The fundamental difference is that the animal visual system does not use a fixed coordinate system to register the positions of the receptive fields. In contrast the computer vision algorithms use the global row and column numbers to label the image pixel positions. The information of these global coordinate numbers becomes baggage for the signal processing algorithm. The invariant features of the visual object are mixed with such baggage which then takes layers of operations to separate if at all possible. The animal vision system must extract the invariant features early on for the quick recognition of deforming objects. The necessity of local coordinate numbers for the neighborhood relations can be met by the feature vectors that encapsulate the neighborhood content without involving a global coordinate system. The eigenvalue vector of the Laplacian matrix formed by the content of the image patch could serve this purpose. However, the eigenvalues are usually mixed with the associated eigenvectors in the spectral decomposition of the Laplacian matrix. These eigenvectors are not invariant even for simple rotations. To separate the eigenvalues and eigenvectors we treat the visual signal on a small receptive field as a linear operator acting on the state of the receptive field. This is in contrast to most spatial signal processing algorithms where the incoming visual signal is treated like an initial condition for the algorithms to act on. We note that in the real biological scenario the visual signal does indeed act like an operator and not like an initial condition.

In this paper we propose a dynamic system model in the above spirit to guide a computer algorithm for such a separation. The invariance acquisition is demonstrated in experiments of deformation transfer. The algorithm works with more general spatial signals such as those representing neuron activities, paving the way for general concept abstraction algorithms that are needed to form nodes in a symbolic AI graph.

The key points are listed below.

- The signal ordering in 2D image and 1D language are fundamentally different. The usual image-wise spatial coordinates for the pixels for 2D image processing unwittingly mix the intrinsic features with the coordinate set up;
- This prevents the processing algorithms from acquiring global-coordinate-independent features for deforming visual objects. Such independence is crucial in forming the matrix representations of the deformations;
- Deformation matrix formation based on actual object deformation images is a primitive case of abstract concept formation. Since the formed matrix is in the same format as the original visual images, further applications of the same procedure could generate the chain of thoughts;
- These matrix representations are nutritious and transparent material for the LLM-like machines to grind on.

II. A "Mental Rotation" Example

In visual image processing one divides the image into patches of pixels. Within each patch the pixel positions are labeled by the row and column numbers within the patch (henceforth refereed as the local coordinates). The order of the patches of the image is labeled by another group of row and column number refereed as the (global) patch

coordinates. Both the local and the patch coordinates cause difficulties in acquiring the invariant features. For example in the convolutional neural network algorithm if the object rotated 90 degrees then all the local and patch coordinates change. The row-major convolution operation cannot co-vary with these changes and result in bad performance. This order issue does not exist in language models since the 1D input data has the temporal order.

We use an MNIST digit orientation transfer example to illustrate our idea for dealing with this order issue. Consider an MNIST digit 8. We would like to make it learn rotation from some training pairs of the MNIST digit 2. Figure 1 shows the results of learning to turn 45 degrees anti-clockwise (center) and 90 degrees clockwise (right) from the upright image in the left. Such "mental rotation" is often used in the
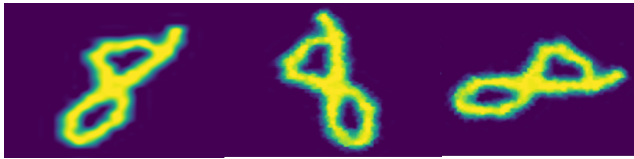

Fig. 1. "Mental Rotation" of a MNIST digit

Raven's progressive matrix IQ test (Figure 2.) Importantly, our algorithm does this learning without the pixel's global row and column numbers. A similar example of a "mental
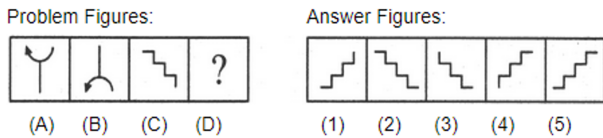

Fig. 2. Non-verbal reasoning test example

Fig. 3. "Mental Affine Transformation" of a MNIST digit

affine transformation" is shown in Figure 3, where the left image is the original, the center is the expected transformed digit 8 and the right is the "mentally transformed" digit 8 as the result of learning the affine transformation from many pairs of digit 2. Figure 4 shows a 2D deformation transfer pipeline. The gist of the algorithm is as follows. We first divide the digit images into small square patches of pixels. For each patch we form a Laplacian matrix. The entry is the absolute difference between any two pixels divided by their distance. We then calculate the eigenvalue vectors of the Laplacian matrices and refer them as sampler vectors. The similarities between the sampler vectors from the original and the deformed images are used to form a mapping matrix. The mapping matrices for the individual training image pairs

are averaged over. The resultant mapping matrix is used to construct the predicted image from the input image. One can see in Figure 1 that the 90 degree rotation prediction is almost perfect, demonstrating the effectiveness of the mapping. The noise in the 45 degree image prediction is due to the mismatch between the rotated image patch and the upright scanning window, and can be reduced by increasing the training pairs or conventional denoising if needed. In
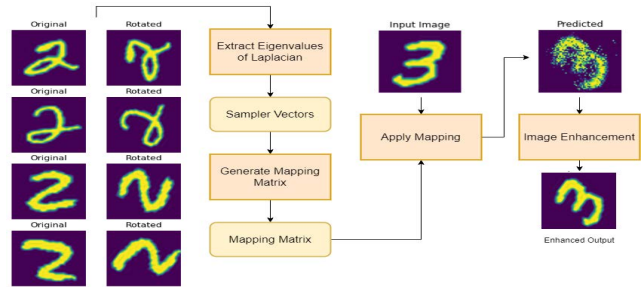

Fig. 4. The 2D Deformation Transfer Pipeline. Given multiple input pairs of images and their corresponding transformed images, the algorithm applies the transformation on a new test image

the Raven's progressive matrix test (Figure 2) there is only one training pair, but it is enough for such simple images. Here each sampler vector of Figure (A) is scanned by all the sampler vectors of Figure (B) to find a sampler vector mapping matrix. This mapping matrix compares the mapping matrices between (C) and all the answer figures to find the best fit. We note that this last step of finding the best fit among the mapping matrices is treating an "abstract concept" representation the same way as treating the "concrete" image signals. In this way the example serves to illustrate how the abstract thinking process is executed in the same algorithms for image processing.

The polar decomposition theorem tells us that a linear transformation can be represented as the product of a rotation or reflection, and a scaling along a set of orthogonal axes. In the animal visual system the individual receptive field is small enough to approximate an object deformation in the field as a linear transformation. The scaling and reflection can be dealt with similarly as the rotations.

We note that in the above orientation transfer example the mapping matrix is the signal representation of an abstract concept that emerged from the training image pairs. In general, a basic challenge for AI is to develop concept abstraction algorithms that can acquire the invariants of objects, whether concrete or abstract, from the instances, to form a signal representation of the concept such as the above mapping matrix. Once we know the representation of the abstract concepts we can test the similarity among them using the same algorithm for visual shape similarity testing. In the above example the mapping matrix for a 45 degree rotation can be compared with other rotation mapping matrices like in the Raven's test. The topic of analogical reasoning for relational similarity testing among ostensibly different scenarios can be cast in this set up.

Figure 4 illustrates the pipeline of the experiment. We first

try to learn the transformation given multiple training pairs of the original images and their rotated counterparts. The concept or transformation is then stored in a mapping matrix. Once the mapping matrix is formed, it can be applied to any number of test images. We note that our experiments are for simple images and transforms due to the constraints of computing resources. The principle algorithm could produce much better illustrations. Here we present one more such experiment in the following Figure 5, where the mapping matrix is learned from a pair of cat images and is applied to a dog image to produce the similarly rotated dog image.
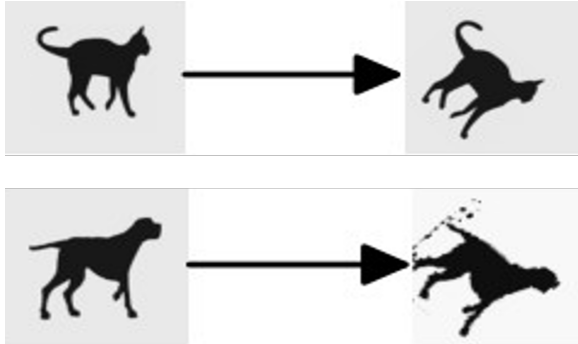

Fig. 5. A Cat rotates 45 degrees and a Dog follows suit

We now discuss how to form the spectral feature vectors, referred as the sampler vectors, from an image. The discussion here is tailored for the current application and is shown in Figure 6. In short, the input image is divided into patches, simulating the multiple receptive fields on the retina. Then, the Laplacian matrix (referred in the Figure as (LBM) for Laplace-Beltrami Matrix for more general discussions) is calculated for each patch. Finally, we calculate the eigenvalues of the Laplacian matrix. In our current example the set of eigenvalues for each Laplace matrix is called the sampler vector, and we use these sampler vectors as features. From the results of our experiments it is clear that the sampler vectors do indeed encapsulate the content of each patch and this encapsulation is independent of the global coordinate numbers and orientation of the patch. Once
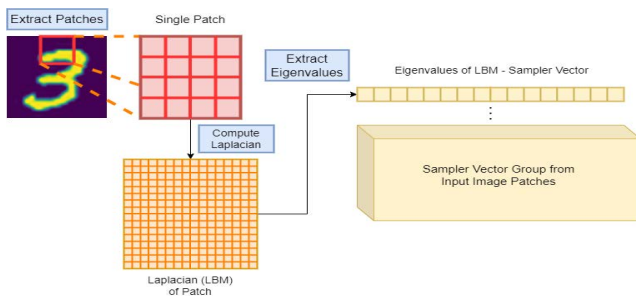

Fig. 6. The Sampler Vector computation. First, we extract multiple patches from the input image. We then compute the Laplacian of each patch and then calculate the eigenvalues of each Laplacian matrix, generating an eigenvalue vector: the Sampler Vector for each patch.

the sampler vectors are generated we compute the cosine similarity between the sampler vectors of the original image

and the transformed image to find the matching patch pairs. The cosine similarities of the matching pairs are used to form a binary mapping matrix to predict the patch positions for constructing the transformed test image. This works thanks to the invariance of the eigenvalue vector against the image rotation. A noteworthy point here is that such invariance is achieved by the artificial order arrangement of the eigenvalue vector components, which has been made into an implicit universal convention in textbooks and numerical packages. Nature would have to use other ways.

Generating a binary mapping matrix from the similarity matrix is not as trivial as taking the argmax along each row. This is because simply taking the argmax along the rows could map multiple patches of the transformed image to the same patch in the original image. To handle this we calculate the argmax row-wise in the order of increasing entropy. This way, the mappings for the most confident rows will be finalized first. For later rows, if the argmax is a particular column which has already been chosen for a previous row then the next best mapping is chosen.

After generating the binary mapping matrix, the mapping is applied to the patches of the new test image in the order of decreasing entropy. This way, badly mapped patches will be overlapped by better patches. Figures 1 and 3 are examples. This will be further discussed in Section IV.

III. A DYNAMIC SYSTEM FOR INVARIANTS SEPARATION

In the orientation transfer example above, the distance between two pixels in the image remains the same. In the affine transform example such distances do not change much. But the method of matching the sampler vectors can be applied more generally. For example, if the deformation is a scaling then the matching sampler vectors will live in pixel patches with correspondingly scaled sizes. This demands the sampler vectors be generated for multiple scales, a task that Nature could accomplish with the receptive field additivity. For image patches there is no such additivity and we need to calculate the sampler vectors at multiple scales.

For general cases it is useful to consider continuous time dynamic models to enlighten discrete computer algorithms. Before diving into the details of the models we highlight some useful thoughts:

- In the brain's neural networks the visual signal elements (pixels) cannot maintain the travel timing synchronization. The spatial signal needs to be converted to temporal signals for transmission.
- In biological brains the similarity testing for spatial signals is hard. On the other hand the temporal signal similarity is easy to test via resonance of signal components.
- Visual spatial signals need to be treated as operators to acquire deformation invariants.
- To model the saccadic motions of the eyes the training images should be augmented with small changes. Invariants of an object are a voting result from the varying images of the object.

We now discuss these in more detail.

A. Treating visual signal as an operator

Conventional object recognition and tracking treat the image matrix as a passive signal source waiting to be processed. But in animal visual systems the light signal is actually an operator acting on the sensor panel. Since in the brain the neural circuit of surrounding-center receptive field automatically executes a Laplace action, the image matrix should be converted to a Laplace matrix. This Laplace matrix of the image is more analogous to the system matrix A of a linear system $\dot{x}(t) = Ax(t)$ than to the initial state $x(0)$, where the vector $x(t)$ is the sensor panel state. If the image matrix has a spectral decomposition $A = \sum_{k=1}^{n} \lambda_k \phi_k \phi_k^{\mathsf{T}}$ then the state $x(t) = \sum_{k=1}^{n} e^{\lambda_k t} \phi_k \phi_k^{\mathsf{T}} x(0)$. One can see that the eigenvalues λ_k and the eigenvectors ϕ_k play separate roles. This has desirable implications. For example, the eigenvalue vector $\lambda = [\lambda_1, \cdots, \lambda_N]$ of the signal matrix A can be arranged as a permutation invariant vector. When A is the Laplace matrix of a small image patch, the vector λ could be used as a local feature encapsulating the patch content and is a simple version of the "sampler vector" of the patch. In this way an object could be represented as a set of such eigenvalue vectors. When two such sets match well we can infer they are probably the same object. Note again that this process is independent of the global coordinate system used to label the image pixel positions.

Note also that in real visual systems the additivity of the receptive field provides sampler vectors for multiple resolutions to enable "cross resolution" mapping for general deformations.

B. Sampler vectors for matrix images

We first restrict to matrix image deformations that can be modeled by the pixel position changes. Consider an image matrix I with pixel (i, j) at the ith row and jth column. When a deformation happens we have $(i, j) \rightarrow (i', j')$, which implies that the corresponding Laplacian matrix L becomes L':

$$L' = PLP^{-1}$$

where the permutation matrix P permutes the ij row and column to $i'j'$ row and column. The eigenvalues of L, $\lambda_k, k = 1, \cdots, N$, arranged in the conventional ascending order, are invariant to this permutation. The eigenvector permutes its components as $\phi_k = P\phi_k, k = 1, \cdots, N$. We note the component sum of the eigenvectors $s_l = \sum_l \phi_k(l)$ is also invariant. The sampler vector of I is a column vector

$$v_k(I) = [\lambda_1, \cdots, \lambda_N, s_1, \cdots, s_N]^{\mathsf{T}}.$$

For small image patches the eigenvector sum component of the sampler vectors can be omitted for computational convenience. These sums are similar to the phase shifts of the Fourier basis in Fourier analysis. For small image patches these are not as important as the eigenvalues which correspond to the component magnitudes in the spectral decomposition of L.

The sampler vectors $v_1, \cdots, v_N$ reconstruct the Laplacian matrix L up to a similarity transformation. An object consists of many small regions. Reconstructions of each small region would be enough for recognizing the object. Now consider a linear differential equation driven by the image matrix I with a uniform initial condition $\psi(x, 0)$

$$\dot{\psi}(x, t) = L\psi(x, t).$$

Suppose that L has the spectral decomposition

$$L = \sum_k \lambda_k \phi_k \phi_k^{\mathsf{T}}.$$

We have

$$\psi(x, t) = \sum_k e^{\lambda_k t} \phi_k(x) \phi_k^{\mathsf{T}}(y) \psi(y, 0).$$

If we sum up the state function $\psi(x, t)$ over a region D of x and call it $h(t)$ we have (assuming that $\psi(x, 0) = 1 \ \forall x$)

$$
\begin{aligned}
h(t) &= \sum_x \psi(x, t) \\
&= \sum_k e^{\lambda_k t} \sum_x \phi_k(x) \phi_k^{\mathsf{T}}(y) \psi(y, 0) \\
&= \sum_k e^{\lambda_k t} \left(\sum_x \phi_k(x) \right) \left(\phi_k^{\mathsf{T}}(y) \psi(y, 0) \right) \\
&= \sum_k e^{\lambda_k t} s_k^2
\end{aligned}
$$

We see that the sampler vector completely determines $h(t)$ and that the time function $h(t)$ reconstructs the system matrix L up to a similarity transform from the expansion

$$h(t) = \sum_n \sum_x e^{Lt} \psi(x, 0) = \sum_n \frac{1}{n!} \sum_x (Lt)^n.$$

By comparing the values of $h(t)$ in the first formula where $h(t)$ is determined by the sampler vector components to the second formula where $h(t)$ is determined by the matrix L and its powers, one can see that under mild conditions the sampler vector determines the entries of the matrix L. We note that this conversion between the temporal and the spatial representations of the same information can be useful for parallel processing of sequential data.

The component magnitude λ_k in the spectral decomposition of L is reflected as the frequency of the "carrier" time function $e^{\lambda_k t}$. As mentioned, the phase information is not important for small regions and the coefficient s_k^2 can be taken as 1, similar to the spectrogram representation of speech signals.

C. The sampler equation

Note that in $h(t)$ the component sum of the k_{th} eigenvector is "carried" by the corresponding time function $e^{\lambda_k t}$ to provide the information of the sampler vector for the downstream receivers. The carrier signals $e^{\lambda_i t}$ are decayed exponential functions of time and are hard to differentiate by the receivers. Second order temporal dynamics would lead to oscillatory carrier signals and the resultant sampler vector would be much more distinguishable by resonators. This

motivates the following sampler equation in the continuous spatial domain:

$$\psi_{tt}(x,t) + \gamma\psi_t(x,t) = (\alpha\Delta + V(x))\,\psi(x,t), \quad (1)$$

$$h(t) = \int_{D_s} \psi(x,t)dx. \quad (2)$$

Here $\psi(x,t)$ is a function of the location vector x and time t. $V(x)$ is the image and γ, α are real constants. D_s is a small region for the sampler to gather information of the spatial signal with the Laplacian Δ. This sampler converts the spatial signal $V(x)$ into an oscillatory temporal signal $h(t)$ for communication and similarity testing. We note that $V(x)$ can be a time varying $V(x,t)$ to model the eye and image movements. The equations are based on physical mechanisms. We treat the light signal of the image as an operator acting on the retina and the center-surround receptive field as a Laplacian. One can also imagine that the memory network is an array of resonators connected to the samplers. These resonators sift the time signals from the visual signal receptive fields according to the resonator frequencies.

We also note that according to the Lie product formula

$$e^{A+B} = \lim_{n\to\infty}(e^{A/n}e^{B/n})^n$$

the two operators in $(\alpha\Delta + V(x))$ can be understood as alternatively acting on the resultant state of the previous steps over very small time intervals, leading to the effect that they are acting simultaneously. In computer algorithms we approximate $(\alpha\Delta + V(x))$ with the Laplacian matrix of the image patch.

Signal conversion via the above equations can be seen with separation of variables for the dynamic state $\psi(x,t)$:

$$\psi(x,t) = X(x)T(t). \quad (3)$$

Substituting into (1) one has :

$$X(T'' + \gamma T') = (\alpha\Delta + V(x))XT. \quad (4)$$

If $T'' + \gamma T' = \lambda T$ and $(\alpha\Delta + V(x))X = \lambda X$ then the equation is satisfied. The equation for the spatial variable $X(x)$ looks for the spectral decomposition of the operator $(\alpha\Delta + V(x))$:

$$(\alpha\Delta + V(x))X = \lambda X. \quad (5)$$

The time variable $T(t)$ obeys

$$T'' + \gamma T' = \lambda T \quad (6)$$

whose solution is an exponentially decayed sinusoidal function (EDS). The spatial eigenvalue λ and the parameter γ jointly determine the frequency of the temporal sinusoid. The output time function $h_{D_s}(t)$ is then the sum of many terms where each term has an EDS signal with the above frequency. The coefficients and the frequencies of these terms are the components of the sampler vector. They encapsulate the content of the visual signal in the sampling domain D_s. Indeed, the realization theory of linear dynamic systems confirms this. If we approximate $\alpha\Delta + V(x)$ by a large

matrix and double the state space to make the equation first order in time we have a first order linear system like $\dot{x}(t) = Ax(t), y(t) = Cx(t)$. The output time function "realizes" the system matrix up to a similarity transform.

D. Invariance of sampler vectors against image deformation

Use of spectral invariants has a long history in computer vision. However most of the work uses a global coordinate system to set up the analysis.

Consider a square image $f(\mathbf{x})$ with the coordinate grid $\mathbf{x}$ and an invertible linear transform A. We have

$$\mathcal{F}f(A\mathbf{x}) = \frac{1}{|\det A|}\mathcal{F}(f)(A^{-\intercal}\xi)$$

where $\mathcal{F}$ denotes the Fourier transform operator, $\mathcal{F}(f)$ denotes the frequency domain function resulted from the Fourier transform acting on the spatial function f, and ξ is the Fourier frequency. The scaling factor $|\det A|$ can be removed with pixel size changes and/or frequency channel scanning. Apart from this scaling factor, the set of functions $\mathcal{F}f(A\mathbf{x})$ is invariant against the change of A in the sense that if $A \to \tilde{A}$, then one can find the values of $\mathcal{F}(f)(A^{-\intercal}\xi)$ at another place $(\tilde{A}^{-\intercal}\xi)$. In other words, if we sample the function $\mathcal{F}(f)$ to form a discrete approximation vector, then in the transformed image we can find a component-wise equivalent vector at the frequencies $\xi'_1, \xi'_2, \cdots, \xi'_n$ that are related to the original frequencies $\xi_1, \xi_2, \cdots, \xi_n$ according to the transform. Consequently, if one can scan the whole frequency range for similarities between the two sets of vectors, and if each set is consisted of mostly distinguishable vectors, then the change of $A \to \tilde{A}$ would not affect the recognition of the original spatial function f.

It might be more direct to see this from

$$\mathcal{F}(\xi) = \int f(\mathbf{x})e^{-i\xi^\intercal \mathbf{x}}dx,$$

$$\mathcal{F}(A^{-\intercal}\xi) = \int f(A\mathbf{x})e^{-i(A^{-\intercal}\xi)^\intercal A\mathbf{x}}d(A\mathbf{x}).$$

The above "global" Fourier analysis does not solve the invariance problem in real vision systems. In a global Fourier transform any local change would affect the transformed values everywhere. In Nature or in computer algorithms one uses small receptive fields or image patches to localize the signal processing. In our setting we use the sampler vectors, which are formed from the spectral analysis of the sampler patches. These patches could be combined at different sizes to reflect the various resolutions of analysis.

In Figure 7 we schematically illustrate a grid of samplers where a curved edge is sampled with the Laplacian samplers shown in yellow. This group of samplers would generate the invariants since they are "covarying" with the edge image. Note that the position labels are based on a pre-fixed global coordinate system which does not matter to us. Only the relative positions of the excited samplers should be in the analysis and the algorithm.

This model reflects the physical process of the visual signal processing. The Laplacian sampler has been acknowledged to model the center-surround receptive fields. In the

continuous model it is additive when covering larger visual fields. In the simplified case of the edge image in Figure 7, the number of Laplacian samplers would covary with the deformed edge length. In the 2D case the samplers would sample all light signals, not just the edges, but the covarying mechanism is the same.
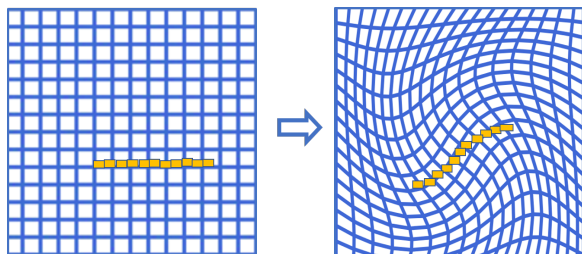

Fig. 7.   Laplacian sampler sets co-vary with the visual scene deformation

We now describe the invariants acquired in the output time function of the sampler. The sampler equation for the initial visual scene $V(x)$ is:

$$\psi_{tt}(x,t) + \gamma\psi_t(x,t) = \left(\alpha\Delta_{xx} + V(x)\right)\psi(x,t), \qquad (7)$$

$$h_{D_s}(t) = \int_{D_s} \psi(x,t)dx, \qquad (8)$$

where $V(x)$ is the spatial signal, $h(t)$ the temporal signal, and D_s the sampling area.

The sampler equation for the deformed visual scene $U(\tilde{x}) = V(x)$ (pixel value at x moves to at $\tilde{x}$) is (with $\tilde{\Delta}$ representing the Laplace-Beltrami operator on a manifold [7] at $\tilde{x}$ region):

$$\psi_{tt}(\tilde{x},t) + \gamma\psi_t(\tilde{x},t) = \left(\alpha\tilde{\Delta}_{\tilde{x}\tilde{x}} + U(\tilde{x})\right)\psi(\tilde{x},t), \qquad (9)$$

$$\tilde{h}(t) = \int_{D_{\tilde{s}}} \psi(\tilde{x},t)d\tilde{x}. \qquad (10)$$

The "Helmholtz equations" and the sampler output functions for the spetral analysis of the original and the deformed images are respectively:

$$\left(\alpha\Delta_{xx} + V(x)\right)\phi_k(x) = \lambda_k\phi_k(x), \qquad (11)$$

$$h(t) = \sum_k T_k(t) \int_{D_s} \phi_k(x)d(x), \qquad (12)$$

and

$$\left(\alpha\tilde{\Delta}_{\tilde{x}\tilde{x}} + U(\tilde{x})\right)\chi_l(\tilde{x}) = \mu_l\chi_l(\tilde{x}), \qquad (13)$$

$$\tilde{h}(t) = \sum_l T_l(t) \int_{D_{\tilde{s}}} \chi_l(\tilde{x})d\tilde{x}. \qquad (14)$$

In the visual systems, the samplers are very dense with small sampler regions and therefore allow linear models. As illustrated in Figure 7, the set of active samplers adapts to changes in the scene. That is, the sampling is covariant with spatial deformations $U(\tilde{x}) = V(x)$. The two output functions have the same "carrier" frequencies thanks to the covariance of the "coordinate grid" (including the sampler domina D_*) and the visual signal. In other words the "carrier" time functions $T_*(t)$ only differ on the "carried" coefficients. The

sampler vector consisting of the "carrier" frequencies and the eigenvector sums are invariant. For small sampler domains we can ignore the eigenvector sums.

IV. MULTIPLE PAIRS OF TRAINING SAMPLES IMPROVE DEFORMATION TRANSFER

Consider a vision task to transfer the deformation from one object (source) to another (target). One can view this deformation acquisition as a primitive case of concept formation. In a real vision situation the images of the source object are always subject to small variations due to small motions of the object, the environment, and the eyes. The deformation we consider is a relatively large motion such as a 45 degree rotation. We try to understand how to form the concept of, say, a 45 degree rotation and apply it to another object. We presented a demo case in Section II.

Assuming that we have n noisy image pairs of an object and its deformation. If we simply take the average of the noisy images of the source object before the deformation and after the deformation, and then proceed with calculations to determine the deformation, we would not get good result. This is primarily because the averaging of the source object image could easily destroy the image features of the object.

Instead we can average the similarities of sampler vectors for denoising. We use data augmentation to generate multiple pairs of the source and target object and then average the sampler vector similarities for patches with the same the origins. The source-target object pairs should be "covarying" during the data augmentation. Note that this is to mimic Nature who uses eye movements for accurate recognition.

V. RELEVANCE TO HUMAN COGNITION PUZZLES

A. Small number sense

In [1] Francois Chollet discusses a benchmark of tasks for machine intelligence challenge, the Abstraction and Reasoning Corpus (ARC), "built upon an explicit set of priors designed to be as close as possible to innate human priors". One of the examples is shown below in Figure 8. The task is to figure out the rule in the three examples shown in the left, and use the rule to solve the challenge on the right. The solution rule is to pick the shape with the most repetitions. This requires a machine algorithm to recognize the numbers of the same shapes in the images. Many animals are capable to sense the difference among the small numbers from one to five. It is far from clear how they do that. In fact, Albert Einstein had used small number sense as an example of concept emergence. "'Three trees' is something different from 'two trees'. Again 'two trees' is different from 'two stones'. The concepts of the pure numbers 2,3,4,..., freed from the objects from which they arose, are creations of the thinking mind which describe the reality of our world."

Research literature has confirmed this. The Wikipedia article on parallel individuation system said the following. "The evidence for parallel individuation system comes from a number of experiments on adults, infants and non-human animals. " "Parallel individuation system in animals was
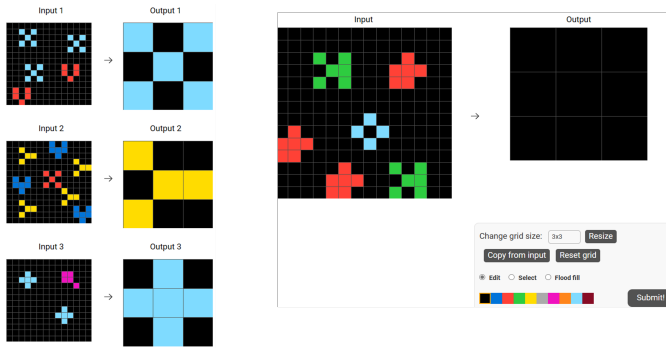
Fig. 8. An ARC task that needs small number sense

demonstrated in an experiment in which guppies were tested on their preference of social groups of different size, under the assumption that they have a preference for bigger size groups. In this experiment, fish successfully discriminated between numbers from 1 to 4 but after this number their performance decreased."

The sampler equation (1) could provide an algorithmic explanation of this. Here the receptive field D_s is the entire visual field. This large D_s is available since the integration is the sum of the parts. The Laplacian matrix formed from an image with three distinct objects is a block diagonal matrix with three blocks. Matrix theory tells us that for such a Laplacian matrix, the multiplicity of the zero eigenvalue equals to the number of blocks regardless of the block sizes. Due to the image background, the Laplacian matrix has non-zero small numbers as the off-block entries. Consequently the output time function for the image that has three objects will contain three close-to-zero frequency components that are separated from the other eigenvalues. Therefore the visual scenes contain 2, 3 and 4 objects would form distinct neuron clusters. The similarity testing algorithm using the sampler vectors for such clusters would form the concepts of 2,3,4. As noted by Burr and Ross (2008): "We propose that just as we have a direct visual sense of the reddishness of half a dozen ripe cherries, so we do of their sixishness. In other words there are distinct qualia for numerosity, as there are for color, brightness, and contrast."

Further more, repeated appearances of the changes from two objects to three objects would allow a brain to form the relational matrix between 2 and 3. The process is to fire the neural clusters representing 2 and 3 together. Then the Spike Timing Dependent Plasticity (STDP) would form a cluster representing the concept "2 to 3". This cluster is also formed for 3 to 4. Similarity of the clusters "2 to 3" and "3 to 4" would form a new cluster representing the concept of "plus one." Now the brain has enough material to form the thought of applying this cluster to "4" and call the result "5". This is a "pure mechanical" explanation of the invention of integers. This iterative process of forming abstract concepts using similarity testing of sampler vector sets generated from neuron clusters could continue further to develop arithmetics and advanced logic systems. As Leopold Kronecker put it:

"God made the integers; all else is the work of man."

The importance of understanding the mechanical and algorithmic ways of generating concepts is to confirm that higher intelligence is essentially the same as lower intelligence. As Steven Pinker [4][5] puts it: "We are made to be gatherers and hunters not scientists"."The genius creates good ideas because we all create good ideas; that is what our combinatorial, adapted minds are for."
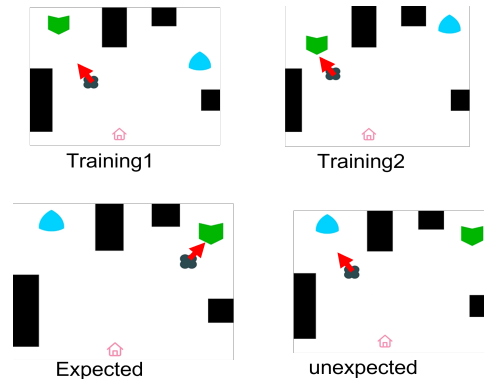
B. Commonsense psychology example

Fig. 9. Commonsense psychology experiment

In a recent paper [2], the authors argued that infants' common sense psychology are foundational to human social intelligence and could thus inform better common sense AI. They experimented with infants and state-of-art AI algorithms and showed that infants are able to pick up the agent intentions while the AI algorithms failed. In Figure 9 the upper row are snapshots of two out of the eight training videos. During training the agent (gray object with red arrow added by us to indicate the motion direction in the videos) moved from the house to the target (green object). Then the infants were shown the bottom row videos. The infants were surprised at the unexpected move of the agent going to the non-target (blue object) while not surprised when the agent is shown to move to the target (green object) despite the location swapping.

This scenario is quite similar to our mapping matrix acquisition experiments. Firstly, in our algorithm there is no global coordinate system to register the object location. Disregarding the global location of objects is a trivial consequence. Secondly, we can form the mapping matrices based on snapshots of the motion videos. The similarities and differences among these mapping matrices would then lead to reactions consistent with the experiment that the intentions are picked up during the training phase.

VI. TOWARDS RELATIONAL CONCEPT GENERATION

The algorithm of mapping matrix generation is intended as a prototype for general concept formation. The recent discussions about the LLMs are very related. Steven Pinker, a leading scholar in human language and psychology, commented that "There are patterns of patterns of patterns of

patterns in the data that we humans can't fathom. It's impressive how LLMs can generate plausible prose, relevant and well-structured, without any understanding of the world — without overt goals, explicitly represented facts, or the other things we might have thought were necessary to generate intelligent-sounding prose. And this appearance of competence makes its blunders all the more striking"[6]. Stephen Wolfram, a famous scientist and an expert in the computational intelligence, commented that the success of LLMs is suggesting there to be major new "laws of language"— and effectively "laws of thought"— out there to discover and that "if we could somehow make the laws explicit, there's the potential to do the kinds of things LLMs does in vastly more direct, efficient—and transparent—ways"[10].

Both experts have emphasized the importance of realizing the "hidden patterns/rules" in human language and human thought processes. The formation of these hidden patterns have "purely mechanical explanation" [8]. The mapping matrix in our experiments shows a prototype. The mapping matrices learned from the deformation pairs of images are themselves matrices and can be treated as images. The algorithm for learning the "first level" mapping matrix between images could be applied to pairs of these first level mapping matrices and generate some "second level" mapping matrices. Figure 10 illustrates this idea. In the Figure we

could form various compositions as thoughts. To this end we note that Alan Turing discussed about the "purely mechanical explanation" of analogical thinking in an interview on October 27, 1949 [8], since it could be achieved with the similarity testing of the mapping matrices. The current large language models achieve some of these capabilities by forming weight groups reflecting the more frequent patterns (See the figurative illustration in Figure 11). This is accomplished with minimizing the next token error. However, as noted, **similarity with no objective is the source of universality and creativity**.
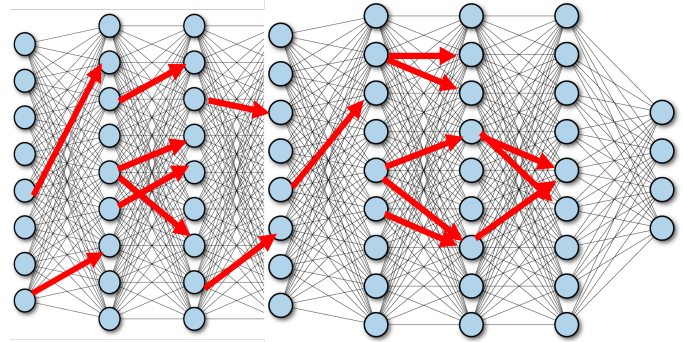
Fig. 11.   Weight groups reflecting more frequent patterns

VII. CONCLUSIONS

Lastly we recall the famous quote of Plato that "Forms (abstract objects) possess a higher degree of reality than do things in the world, which are changeable and always coming into or going out of existence." We could apply this principle to neural networks where the more abstract objects are easier to reflect in some weight groups since they are more frequently repeated in the neural network weight-adjusting processes. Our mapping matrix extraction algorithm provides some such objects since the mapping matrices are representations of relations. See also [3].
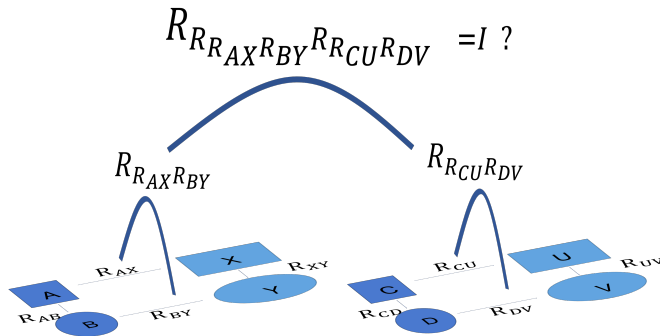

Fig. 10.   A Tower of Relational Concepts

use capital "R" to indicate that the mapping matrices are relational matrices. The bottom blocks marked by A, B, X, Y, C, D, U, V are images. R_{AX} is the mapping matrix between the image pair A and X, and so on. The top equation is an example to check if the mapping $R_{R_{CU}R_{DV}}$ is an inverse of the mapping $R_{R_{AX}R_{BY}}$. For example, if R_{AX} and R_{BY} are 20 and 50 degree clockwise rotations, R_{CU} and R_{DV} are 20 and 50 degree anti-clockwise rotations. Then $R_{R_{AX}R_{BY}}$ and $R_{R_{CU}R_{DV}}$ are clockwise and anti-clockwise rotation of 30 degree respectively. The equation would hold since these two rotations would move the pixels back to their original positions. This could tell if the right side transformation compositions would give an inverse to the left side transformations.

We believe this explains an aspect of the emergence of logic reasoning capability in animal brains. When many such mapping matrices accumulate in the brain circuits they

REFERENCES

[1] F. Chollet. On the measure of intelligence. *CoRR*, abs/1911.01547, 2019.
[2] G. Stojnic et.al. Commonsense psychology in human infants and machines. *Cognition, Elsevier*, 235, 2023.
[3] M. Mitchell. Why the abstraction and reasoning corpus is interesting and important for ai. *substack*, Feburary 28, 2023.
[4] S. Pinker. *The Stuff of Thought*. Viking Penguin, 2007.
[5] S. Pinker. The cognitive niche: Coevolution of intelligence,sociality, and language. *PNAS*, 107(May):8993–8999, 2010.
[6] S. Pinker and A. Powell. Will chatgpt supplant us as writers, thinkers? *The Harvard Gazette*, Feburary 14, 2023.
[7] S. Rosenberg. *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*. London Mathematical Society Student Texts. Cambridge University Press, 1997.
[8] A. Turing. A lecture and two radio broadcasts on machine intelligence by alan turing. *Machine Intelligence*, 15, 1999.
[9] S.M. Ulam. *Adventures of a Mathematician, The quote appears on page 222 of the 1991 edition of the book*. University of California Press, 1976.
[10] S. Wolfram. What is chatgpt doing . . . and why does it work? *Stephen Wolfram Writings*, Feburary 14, 2023.