Artificial neural network modelling of the neural population code underlying mathematical operations

Tomoya Nakai\textsuperscript{a,b,c}, Shinji Nishimoto\textsuperscript{a,c,d}

\textsuperscript{a} Center for Information and Neural Networks, National Institute of Information and Communications Technology, Suita, Japan
\textsuperscript{b} Lyon Neuroscience Research Center (CRNL), INSERM U1028 - CNRS UMR5292, University of Lyon, Bron, France
\textsuperscript{c} Graduate School of Frontier Biosciences, Osaka University, Suita, Japan
\textsuperscript{d} Graduate School of Medicine, Osaka University, Suita, Japan


corresponding author at: Centre Hospitalier Le Vinatier (Bât. 452), 95 Bd Pinel, 69500 Bron, France.
E-mail address: nakai.tomoya@neuro.mimoza.jp (T. Nakai).

https://doi.org/10.1016/j.neuroimage.2023.119980.
Received 5 November 2022; Received in revised form 10 February 2023; Accepted 23 February 2023
Available online 26 February 2023.

\textsuperscript{7} Corresponding author at: Centre Hospitalier Le Vinatier (Bât. 452), 95 Bd Pinel, 69500 Bron, France.
E-mail address: nakai.tomoya@neuro.mimoza.jp (T. Nakai).

1. Introduction

Mathematical operations have long been regarded as a sparse, symbolic (or categorical) process in human neuroimaging studies. For example, neuroimaging data under subtraction and multiplication conditions was analysed based on sparse event-related designs including two operators (Ischebeck et al., 2006; Prado et al., 2011). In contrast, recent advances in artificial neural networks (ANN) enable extracting distributed representations of quantity information (Kim et al., 2021; Nasr et al., 2019) and mathematical operations (Russin et al., 2021; Schlag et al., 2019), where each operator is represented as a vector in a high-dimensional latent feature space. In particular, a recent study demonstrated that the intermediate layers of the tensor-product (TP)-transformer model can capture the compositional structure of mathematical expressions (Russin et al., 2021).

An emerging field of cognitive computational neuroscience has reported evidence on whether such latent features can explain the symbolic representations in the human brain. Previous studies have demonstrated that ANNs and biological neural networks (BNNs) partially share representations for visual (Groen et al., 2018; Güçlü and van Gerven, 2015; Horikawa and Kamitani, 2017; Kietzmann et al., 2019; Yamins et al., 2014), auditory (Kell et al., 2018; Kounura et al., 2019) and language processing (Caucheteux and King, 2022; Goldstein et al., 2022; Hannagan et al., 2021; Schmitt et al., 2021; Schrimpf et al., 2021). However, the representational relationship between ANN and BNN in mathematics has not yet been investigated.

Previous studies have used the encoding model approach (Naselaris et al., 2011) to compare the representational relationship between ANNs and BNNs (Caucheteux and King, 2022; Goldstein et al., 2022; Jain and Huth, 2018; Ratan Murty et al., 2021; Schmitt et al., 2021; Schrimpf et al., 2021). Encoding models quantitatively predict brain activity based on a combination of features extracted from the presented stimuli. Researchers have adopted this approach to comprehensively examine visual (Kay et al., 2008b; Naselaris et al., 2009; Nishimoto et al., 2011), auditory (Nakai et al., 2021; Norman-Haignere et al., 2015), semantic (Huth et al., 2016, 2012; Popham et al., 2021), emotional (Horikawa et al., 2020; Koide-Majima et al., 2020) and many complex cognitive functions (Nakai and Nishimoto, 2020).
Encoding models also allow the comparison between sparse and latent features in the same brain data using representational similarity analysis (RSA) and feature-brain similarity (FBS) (Nakai et al., 2021). The RSA is a region of interest (ROI)-based approach. In contrast, the FBS can directly assess the relationship between latent features and brain representations in each voxel and how such a relationship produces brain activation patterns of mathematical operations modelled by sparse symbolic features.

To investigate the relationship between the symbolic and distributed representations underlying mathematical operations in the brain, we used part of a three-hour fMRI dataset (Nakai and Nishimoto, 2023) where subjects performed a series of mathematical problems with nine different combinations of operators (Fig. 1A). This dataset is characterised by a relatively large sample size (N = 3255) per individual, in contrast to existing mathematical fMRI datasets (e.g. N = 430 in Suárez-Pellicioni et al. (2019)). By using sparse operator features, we constructed a voxel-wise encoding model to predict the brain activity specific to each mathematical operation (Fig. 1B). We then extracted the latent features from the intermediate layers of an ANN (TP-Transformer) (Schlag et al., 2019) and constructed another encoding model based on these latent features. We hypothesised that ANN-based distributed representations can explain the brain activity patterns underlying mathematical operations depicted by sparse features. Alternatively, such a sparse representation could be explained by other features not specific to mathematics.

2. Methods

2.1. Subjects

Eight healthy college students (aged 20–23 years, three females, all with normal vision), denoted as ID01–ID08, participated in this study. This number is in the same range as previous encoding model studies (Horikawa et al., 2020; Huth et al., 2016; Nakai and Nishimoto, 2020; Pohpan et al., 2021). Subjects were all right-handed (laterality quotient, 80–100) as assessed using the Edinburgh inventory (Oldfield, 1971).

2.2. Stimuli and procedure

We selected arithmetic problems with a single operation, addition (Add), subtraction (Sub), multiplication (Mul) and division (Div), and arithmetic problems with two operations, including addition and subtraction (AddSub), addition and multiplication (AddMul), addition and division (AddDiv), subtraction and multiplication (SubMul) and subtraction and division (SubDiv). Each task consisted of 35 instances. Subjects also underwent another math problem-solving condition in word format and a control sentence comprehension condition; these conditions were analysed elsewhere (Nakai and Nishimoto, 2023). Behavioural data were also analysed in this previous work.

In each trial, an arithmetic expression problem (e.g. ‘3 x 2 = ?’) was presented for 4 or 6 s (4 s for single-operator problems, and 6 s for double-operator problems). The subjects were instructed to perform a calculation based on the presented problem and to press the button on the left when they had an answer. The fixation cross stimulus was then presented for 1–2 s, after which a probe digit stimulus was presented for 2 s (e.g. ‘6’). Subjects were again asked to press the left or right button if the presented digit matched or did not match their answer, respectively. The next trial began after the fixation cross was again displayed for 1–2 s. The number of letters in each task was as follows: 5.5 ± 0.5 (Add), 6.5 ± 0.5 (Sub), 5.0 ± 0.2 (Mul), 5.9 ± 0.4 (Div), 9.3 ± 0.9 (AddSub), 8.9 ± 0.3 (AddMul), 10.1 ± 0.7 (AddDiv), 9.6 ± 0.5 (SubMul) and 10.9 ± 0.5 (SubDiv). Stimuli were presented on a projector screen inside the scanner (21.0 × 15.8° visual angle at 30 Hz). During scanning, subjects wore MR-compatible ear tips. The experiment was performed over three days, with six runs per each day. A presentation software (Neurobehavioural Systems, Albany, CA, USA) was used to control stimulus presentation and collection of behavioural data. Optic response pads with two buttons were used to measure button responses (HHSC-2 × 2, Current Designs, Philadelphia, PA, USA).

Six training and two test runs were conducted. The data from each pair of test runs were averaged to increase the signal-to-noise ratio. Each run contained 45 trials consisting of five trials for each of the nine tasks; the presentation order was randomised within the run. A single run lasted 455 s. At the beginning of each run, we acquired 10 s of dummy scans, during which the fixation cross was displayed; these dummy scans were later omitted from the final analysis to reduce noise. We also obtained 10 s of scans at the end of each run, during which the fixation cross was displayed, and which were included in the analyses.

2.3. MRI data acquisition

The experiment was conducted using a 3.0 T scanner (MAGNETOM Prisma; Siemens, Erlangen, Germany) with a 64-channel head coil. We scanned 72 interleaved axial slices, 2-mm thick without a gap, parallel to the anterior and posterior commissure line, using a T2*-weighted gradient-echo multiband echo-planar imaging sequence (repetition time (TR) = 1000 ms; echo time (TE) = 30 ms; flip angle (FA) = 62°; field of view (FOV) = 192 × 192 mm2; resolution = 2 × 2 mm2; multiband factor = 6). We obtained 465 vol for each run, following 10 dummy images. As anatomical reference, high-resolution T1-weighted images of the whole brain were acquired from all subjects with a magnetisation-prepared rapid acquisition gradient-echo sequence.
(TR = 2530 ms; TE = 3.26 ms; FA = 9°; FOV = 256 × 256 mm²; voxel size = 1 × 1 × 1 mm³).

2.4. fMRI data pre-processing

In each run, motion correction was performed using the statistical parametric mapping toolbox (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). All volumes were aligned to the first EPI image for each subject. Low-frequency drift was removed using a median filter with a 120-s window. Slice timing correction was performed against the first slice of each scan. The response for each voxel was then normalised by subtracting the mean response and scaling it to the unit variance. We used FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) to identify the cortical surfaces from anatomical data and register these to functional data voxels. For each subject, identified voxels in the cerebral cortex (59,499–75,980 voxels per subject) were used for analysis.

2.5. Stimulus features

2.5.1. Operator features

Operator features included one-hot vectors with assigned values of 1 or 0 for each time bin during the stimulus presentation of arithmetic problems, indicating whether one of the tested tasks (Add, Sub, Mul, Div, AddSub, AddMul, AddDiv, SubMul and SubDiv) was performed in that period. All trials were assigned with these values, regardless of their RTs or response accuracy. Thus, nine operator features were used.

2.5.2. ANN features

ANN features were extracted using a pretrained TP-transformer model (Schlag et al., 2019). TP-Transformer, an extension of the transformer model (Vaswani et al., 2017) consisting of encoder and decoder layers, where the former layer processes given math expressions and the latter generates the answer. Both encoder and decoder have six transformer layers containing multi-head attention modules with eight heads.

In contrast to the regular transformer where the attention head consists of query, key and value vectors, TP-Transformer additionally uses a role vector to better capture structural information in the input math expressions. Each attention head thus consists of query (Q), key (K), value (V) and roll (R) vectors for each input symbol:

\[ TP - Attention(Q, K, V, R) = \text{softmax} \left( \frac{Q K^T}{\sqrt{d_{\text{model}}}} \right) V \otimes R \]

where \( d_k \) is a dimension of K vector and \( \otimes \) is Hadamard product. The other part of the architecture and hyper-parameters of the TP-Transformer is similar to the regular transformer (dimensionality of input and output, \( d_{\text{model}} = 512 \); dimensionality of feed-forward network, \( d_f = 2048 \)). The model was trained using the Mathematics Dataset (Saxton et al., 2019).

For each input math expression (e.g. ‘3 + 5’), we extracted the average query, key, value and roll vectors from each of the six encoding layers. Time points without vector assignments were defined as 0. Although we could use all these vectors, we used the query vectors of the sixth encoding layer as a representative vector based on a previous study (Russin et al., 2021). We also extracted query vectors of the sixth encoding layer with each mathematical symbol (e.g. ‘3’, ‘+’, etc.) as input. ANN features had 512 dimensions.

2.5.3. ME features

We employed a previously described ME model (Koide-Majima et al., 2020; Nakai and Nishimoto, 2020; Nishimoto et al., 2011) available in a public repository (https://github.com/gallantlab/motion_energy_matlab). First, movie frames and pictures were spatially downsampled to 96 × 96 pixels. RGB pixel values were then converted into the Commission International de l’Eclairage (CIE) LAB colour space; then, colour information was discarded. The luminance (L’) pattern was passed through a bank of three-dimensional spatiotemporal Gabor wavelet filters; the outputs of two filters with orthogonal phases (quadrature pairs) were squared and summed to yield the local ME. Subsequently, ME was compressed with a log-transformation and temporally downsampled to 0.5 Hz. Filters were tuned to six spatial (0, 1.5, 3.0, 6.0, 12.0 and 24.0 cycles per image) and three temporal frequencies (0, 4.0 and 8.0 Hz) without directional parameters. Filters were positioned on a square grid covering the screen. Adjacent filters were separated by 3.5 standard deviations of their spatial Gaussian envelopes. To reduce the computational load, the original ME features, which had 1395 dimensions, were reduced to 300 dimensions using principal component analysis.

2.5.4. Word2vec features

To quantitatively evaluate the brain representations of the presented semantic information in a data-driven manner, we extracted the semantic features from each narrative stimulus using word2vec (Mikolov et al., 2013). Word2vec, based on the skip-gram algorithm, has been used to embed words into distributed representations. We used a text corpus from the Japanese Wikipedia dump on July 9, 2019 (http://dumps.wikimedia.org/jawiki/) to learn the skip-gram vector space. All math expressions were segmented into words (i.e. symbols) and morphologically parsed using MeCab (https://taku910.github.io/mecab/). Individual word segments were projected into the 300-dimensional space. Timepoints without any word vector assignments were defined as 0. The resultant concatenated vectors were downsampled to 1 Hz. The dimension size of word vectors was set to the default value of 300.

2.5.5. BR feature

The single button response (BR) feature was constructed based on the number of button responses per second.

2.5.6. Letter feature

The single letter feature was constructed based on the number of letters appearing in each stimulus.

2.5.7. RT feature

For each time bin during presentation of mathematical problems, the single trial RT was assigned as the trial’s cognitive load.

2.6. Encoding model fitting

In the encoding model, the cortical activity in each voxel was fitted with a finite impulse response model that captured the slow hemodynamic response and its coupling with neural activity (Kay et al., 2008a; Nishimoto et al., 2011). The feature matrix \( F_E \) \([T \times 6 N]\) was modelled by concatenating sets of \([T \times N]\) feature matrices with six temporal delays between 2 and 7 s \([T \times N]\) was number of samples; \(N = \) number of features). The cortical response \( R_E \) \([T \times V]\) was then modelled by multiplying the feature matrix \( F_E \) by the weight matrix \( W_E \) \([6 N \times V]\) \([V = \) number of voxels]:

\[ R_E = F_E W_E \]

We used L2-regularised linear regression with the training dataset (consisting of 2790 [2790 s] samples) to obtain the weight matrix \( W_E \). The optimal regularisation parameter was assessed using 10-fold cross-validation where the 11 different regularisation parameters ranged from 1 to \(10^\text{10} \).

The test dataset consisted of 465 samples (465 s, repeated twice). Two repetitions of the test dataset were averaged to increase the signal-to-noise ratio. Statistical significance (one-sided) was computed by comparing estimated correlations to the null distribution of correlations between two independent Gaussian random vectors with the same length as the test dataset. The statistical threshold was set at \(p < 0.05\) and corrected for multiple comparisons using FDR (Benjamini and Hochberg, 1995). For data visualisation on cortical maps, we used pycortex (Gao et al., 2015) and fsbrain (Schäfer and Ecker, 2020).
2.7. Encoding model fitting excluding regressors of noninterest

To exclude the possible effect of sensorimotor processing, linguistic processing and general cognitive load on the model predictions, we concatenated ME (visual), word2vec (semantic), BR (motor), RT (cognitive load) and letter (orthographic) features to the original operator and ANN feature matrices. The concatenated features were used as a new feature matrix for modelling encoding. L2-regularised linear regressions were applied as described in the Encoding model fitting subsection. For model testing, prediction accuracy was evaluated using the feature matrix after excluding non-interesting features.

2.8. RSA

To compare ANN features with the sparse operator model, we performed RSA based on two different representational similarity matrices (RSMs). First, we calculated a reference feature vector for each operator type using the average features extracted from TP-Transformer for all trials in the training dataset. A feature-based RSM was calculated using the Pearson correlation distance across all combinations of the reference feature vectors for the nine tasks. Second, for each task, we averaged the weight vectors of the operator encoding model for all voxels included in each target anatomical ROI; six-time delays were further averaged. A brain-based RSM was calculated using the Pearson correlation distance across all combinations of the resultant mean weight vectors for the nine tasks. The upper triangular parts of the feature-based and brain-based RSMs were rearranged into single vectors and the Spearman’s correlation coefficients between the two vectors were calculated.

2.9. FBS analysis

To provide voxel-wise similarity information and differences in patterns across the nine tasks, we calculated the FBS for each cortical voxel by calculating the Pearson’s correlation coefficient between the reference feature vector and the voxel-specific weight vector in each cortical voxel. The voxel-specific weight vector was taken from each target model weight matrix and averaged for six temporal delays.

2.10. Decoding model fitting

In the decoding model, the cortical response matrix $R_D \{ T \times 6 \}$ was modelled using concatenating sets of $\{ T \times V \}$ matrices with temporal delays of 2–7 s. The feature matrix $F_D \{ T \times N \}$ was modelled by multiplying the cortical response matrix $R_D$ by the weight matrix $W_D \{ 6 V \times N \}$:

$$F_D = R_D W_D$$

The weight matrix $W_D$ was estimated using an L2-regularised linear regression with the training dataset, following the same procedure as for encoding model fitting.

To examine the generalizability of our models, we performed three types of novel task decoding. Here, we use the term ‘novel’ to indicate decoding analyses where data on target tasks are not used in the training samples. (1) First, we performed nine independent model fittings, each using a different task as a target. From the training dataset, we excluded the time points at which the target task is presented and those within 7 s after its presentation. In the test dataset, we used only the time points at which the target task is presented and those within 7 s after its presentation. This allowed us to assume that the activity induced by the target task (e.g. ‘AddSub’) and that induced by the other eight tasks (‘Add’, ‘SubDiv’, etc.) did not overlap, and to investigate decoding accuracies for novel tasks. (2) Second, we performed two independent model fittings, where only single- or double-operator problems were used as training dataset; the remaining tasks (double-operators in case of single operators used in model training) were used in the model testing phase. (3) Third, we performed four independent model fittings with tasks containing one of the four operators (‘Add’, ‘Sub’, ‘Mul’ and ‘Div’) as targets, where tasks without the target operator were used in the model training phase (e.g. without ‘Add’, ‘AddSub’, ‘AddMul’ and ‘AddDiv’ in the case of ‘Add’ as a target).

The above decoding models outputs a series of 512 dimensional vectors (decoded vectors). For the evaluation of decoding accuracy, we measured the similarity between each of the nine reference feature vectors (see FBS analysis) and the decoded vector for each time point using Pearson’s correlation coefficients. We refer to this correlation coefficient as task score (Nishida and Nishimoto, 2018). We then plotted a receiver operating characteristic (ROC) curve for each target task by plotting the ratio of true and false positives by changing the discrimination threshold (from −1 to 1) based on the task score; that is, only time points where the obtained task score was higher than the discrimination threshold were assigned with this task. Decoding accuracy was evaluated using the area under the ROC curve (AUC). The statistical significance of the decoding accuracy for each task was tested using a one-sided permutation test ($p < 0.05$, with FDR correction).

2.11. Noise ceiling

To quantify the signal-to-noise ratio of test samples, we applied a noise ceiling calculation (Schoppe et al., 2016). This measurement is based on the formula introduced by Sahani and Linden (2002) to split the variance of repetitive neural responses (i.e. total power; TP) into estimators of signal power (SP) and noise power (NP):

$$TP = \frac{1}{N} \sum_n V ar(r_n)$$

$$SP = \frac{1}{N-1} (N \times V ar(\frac{1}{N} \sum_n r_n) - TP)$$

$$NP = TP - SP$$

Where $N$ is the total number of repetitions ($\approx 2$ in the current study), $r_n$ is a time series of neural responses at the $n$th repetition. Based on these estimators, the noise ceiling value $NC$ (maximum prediction accuracy) can be calculated as follows:

$$NC = \sqrt{SP + \frac{NP}{N^2}}$$

The statistical significance of NC was assessed using a null distribution of the same measure obtained based on randomly generated responses (Gaussian random vectors). The statistical threshold was set at $p < 0.05$ and corrected for multiple comparisons using FDR (Benjamini and Hochberg, 1995).

Finally, the original prediction accuracy was corrected using the noise ceiling value in each voxel as follows:

$$R_{\text{corrected}} = \frac{R_{\text{original}}}{NC}$$

Where $R_{\text{original}}$ denotes the original prediction accuracy, and $R_{\text{corrected}}$ denotes the corrected one.

3. Results

3.1. Latent features extracted from ANN predicted the brain activity induced by mathematical problems

To investigate whether the symbolic operator and latent ANN features capture brain representations of math problem solving, we constructed voxel-wise encoding models with both types of features using the training dataset and quantified the prediction accuracy of each model with the test dataset (Fig. 1B). Both models significantly predicted the activity of large brain regions of the bilateral frontal, parietal and occipital cortices (operator model, prediction accuracy across whole cortex $= 0.040 \pm 0.026$ (mean ± s.d.), $21.5 \pm 9.4\%$ of voxels were significant; ANN model, prediction accuracy $= 0.045 \pm 0.025$, $22.8 \pm 10.0\%$ of voxels were significant; Figs. 2A, S1). The overall prediction maps between operator and ANN models were similar across all cortical voxels (Spearman’s correlation coefficient, $\rho = 0.783 \pm 0.088$; Figs. 2B, S2).

Encoding models across the whole cortex likely contain voxels that do not respond to mathematical problems, resulting in a low signal-to-noise ratio and unreliable comparison between the two models. To
address this issue, we calculated noise ceiling (maximum prediction accuracy) using two repeated test runs (Schoppe et al., 2016). Within the voxels showing significant noise ceiling values (0.194 ± 0.033 across whole cortex; Fig. S3), the operator model significantly predicted 38.9% ± 21.7% of voxels (prediction accuracy = 0.092 ± 0.073; corrected prediction accuracy = 0.166 ± 0.130), and the ANN model significantly predicted 39.3% ± 21.6% of voxels (prediction accuracy = 0.097 ± 0.069; corrected prediction accuracy = 0.189 ± 0.140). Moreover, we found a significant correlation of prediction maps between operator and ANN models (p = 0.819 ± 0.115).

The above results suggest that both operator and ANN encoding models equivalently predict brain activity during math problem solving. However, these models may partially contain non-mathematical information. To exclude the influence of non-interesting features, we constructed additional encoding models concatenating the target features with other non-mathematical features such as motion energy (visual), reaction times (procedural load), number of letters (orthographic), word2vec (semantic) and button responses (motor). The operator model significantly predicted 18.3 ± 8.0% of voxels (prediction accuracy = 0.028 ± 0.024) and the ANN model significantly predicted 21.7 ± 10.4% of voxels (prediction accuracy = 0.043 ± 0.026). Once again, the prediction maps between the operator and ANN models were positively correlated (p = 0.667 ± 0.078). A positive correlation between the two models was found as well when using only voxels with significant noise ceiling values (p = 0.718 ± 0.126) (Fig. S4). These results are consistent with the argument that the ANN model is relevant to latent information underlying mathematical operations.

Next, we examined how the information contained in the ANN differs depending on math expression components. We constructed ANN models using features from different positions in math expressions (e.g. ‘4’ and ‘+’ from ‘4 + 3’) and calculated the Spearman’s correlation coefficients between the prediction accuracy maps of the operator model and various ANN models. We found larger correlation coefficients for most subjects when using ANN features from operators for double-operator problems (difference of Spearman’s correlation coefficients, 0.102 ± 0.070; two-tailed Wilcoxon signed-rank test across subjects, p = 0.008) but not for single-operator problems (0.121 ± 0.244, p = 0.148) (Fig. 2C and D). This effect was particularly

Fig. 2. Prediction of brain activity by the ANN and operator encoding models. (A) Prediction accuracies of the sparse operator (red) and latent ANN (green) models shown on the cortical surface of subject ID01. Only voxels with significant prediction are shown (p < 0.05, FDR corrected). (B) Scatter plot of prediction accuracies of ANN and operator models, shown for both the original model (cyan) and after exclusion of non-mathematical regressors (yellow). (C-F) Correlation between the prediction map of the operator model and those of ANN models, constructed using features from different positions, shown for both (C) single- and (D) double-operator problems in the original model and after exclusion of non-mathematical regressors (E, F). Individual subjects’ data are indicated by small circles. Error bar, SD.
evident when non-mathematical features were regressed out (single-operator problems, 0.096 ± 0.059, p = 0.008; double-operator problems, 0.068 ± 0.042, p = 0.008; Fig. 2E and F). These results demonstrate that the ANN model captures representations of different mathematical operations.

3.2. The ANN model captured symbolic representations of different math operations

To interpret the latent features of the ANN model and compare it with the sparse operator model, we performed an RSA (Fig. 3A). In each anatomical ROI, we calculated a RSM with the correlation distance amongst nine tasks using the operator model (brain-based RSM) and another using ANN features (feature-based RSM). We found a significant correlation between the upper triangular matrices of brain-based and ANN feature-based RSMs in the bilateral frontal, parietal, temporal and occipital cortices (Fig. 3B). Amongst cortical ROIs, we observed the largest correlation coefficients in the bilateral intraparietal sulci (IPS; left IPS, ρ = 0.418 ± 0.093; right IPS, ρ = 0.416 ± 0.066).

As control analyses with non-mathematical latent features, we also calculated feature-based RSMs using other latent features of the word2vec (semantic) and motion energy (visual) models. The word2vec is a widely-known model of semantic information used in previous encoding/decoding studies (Nishida et al., 2021; Nishida and Nishimoto, 2018). This model can extract semantic features from each number and operator but does not capture the compositional information arising from combinations of numbers and operators (e.g. ‘8’ from a given expression ‘3 + 5’). The motion energy model is used to reconstruct visual information from brain activity (Nishimoto et al., 2011). The word2vec model showed a left-lateralised pattern and larger correlation coefficients in the left inferior frontal and superior temporal cortices (Fig. 3C), whereas the motion energy model showed the largest correlation coefficients in the bilateral occipital cortex (Fig. 3D).

3.3. Latent ANN features could reconstruct brain representations of the sparse operator model

Although RSA calculates similarity in an abstract space, it does not directly assess whether there is a relationship between latent features and brain representations. Moreover, RSA calculates similarity based on geometric multivoxel patterns and does not provide similarity information for each voxel. To address these issues, we performed an FBS analysis as previously described (Nakai et al., 2021). Specifically, we calculated the Pearson’s correlation coefficient between the reference vector of each operator and the weight vector extracted in each cortical voxel (Fig. 4A).

This way, we obtained a single FBS value for each cortical voxel and for each operator. The resultant FBS map obtained using latent ANN features (Fig. 4B) was similar to the weight map of the sparse operator model (Fig. 4C). We found a significant correlation between the FBS map and the operator-weight map (mean correlation coefficient across nine tasks, 0.724 ± 0.031, p < 0.001: Fig. 4D, S5). By comparing different latent features, we found that ANN-extracted features (TP-transformer) outperformed those extracted from motion energy (visual) and word2vec (semantic) models using whole-cortical voxels (Wilcoxon signed-rank test; p = 0.008) (Fig. 4E). Furthermore, the correlation between the operator-weight map and FBS values increased with ANN hierarchy (difference between encoding layer-6 and other lay-
ers, Wilcoxon signed-rank test, \(p < 0.008; \text{Fig. 4F, S6}\). Similar results were also found when using only voxels with significant noise ceiling values (Fig. S7). These results indicate that the latent ANN features can serve to reconstruct the brain representations of mathematical operations obtained with symbolic features.

3.4. The ANN model enables decoding novel math problems

To further examine the generalizability of the ANN model across different mathematical operations, we performed a series of decoding analyses based on ANN features (Fig. 5A). While encoding models can quantitatively assess differences between brain regions with respect to feature representations, decoding models have an advantage in quantitatively evaluating differences between stimulus features with respect to brain responses. The decoding approach is thus useful to examine whether certain operators are more easily decodable from brain responses and whether their information can be generalised. First, the generalizability of decoding models was evaluated by training a decoding model using eight of the nine tasks in the training dataset (e.g. ‘Add’, ‘Sub’, ‘Mul’, ‘Div’, ‘AddMul’, ‘AddDiv’, ‘SubMul’ and ‘SubDiv’) and testing it with the remaining operator in the test dataset (e.g. ‘AddSub’). Second, decoding accuracy was calculated as the AUC (Fig. S8) (Nishida and Nishimoto, 2018). As a result, mathematical problems were significantly decoded for all subjects (mean decoding accuracy across nine tasks, \(0.724 \pm 0.031\); one-sided permutation test for average decoding accuracy, \(p < 0.001\), FDR corrected; Fig. 5D).

Next, we investigated if the generalizability of decoding accuracy was confined to a specific problem structure (i.e. single- or double-operators), as math problem structure may affect brain activity patterns (Nakai and Sakai, 2014). We thus trained decoders using single operators in the training dataset and tested with double-operators in the test dataset; inversely, decoders trained using double-operator problems were applied to single-operator problems in the test dataset (Fig. 5B). In both cases, the decoding accuracy was larger than chance (0.5) for all subjects (mean decoding accuracy across nine tasks, \(0.670 \pm 0.044\); \(p < 0.001\), FDR corrected; Fig. 5E).

Although the above analysis significantly decoded operators not used for model training, since part of the single operators in the training dataset (‘Add’ and ‘Mul’) were also used amongst the double-operators in the test dataset (e.g. ‘AddMul’), we constructed additional decoding models without a target operator (e.g. without ‘Add’, ‘AddSub’, ‘AddMul’ and ‘AddDiv’) and tested them with the target operator (e.g. ‘Add’) (Fig. 5C). We again found significant decoding accuracy for all four target operators (mean decoding accuracy across nine tasks, \(0.721 \pm 0.034\); \(p < 0.001\), FDR corrected; Fig. 5F). These results indicate that the ANN model captures detailed representations of different mathematical problems.

4. Discussion

In the current study, we constructed voxel-wise encoding models using both a sparse operator and latent ANN features for math problem solving and examined the representational relationship in the brain.
between these two features types. Representational similarity and FBS analyses demonstrated shared representational patterns between ANN and BNN. This effect was particularly evident in the IPS. Latent ANN features further allowed decoding of novel operators not used for model training. These results indicate that ANN-based distributed representations in the brain partially explain the symbolic processing of mathematical operations.

Correlation analyses of prediction maps demonstrated similar prediction patterns between the sparse operator and latent ANN models based on mathematical operations (Fig. 2C, D). These results are consistent with a previous study using the same ANN model (Russin et al., 2021), which showed that the feature vectors of intermediate calculations of a math expression were more similar to those extracted from operators of the same expression than from digits. Although we did not use any control condition, we excluded the possible effects of non-mathematical features by concatenating motion energy (visual), reaction times (procedural load), number of letters (orthographic), word2vec (semantic), and button responses (motor) features to the original feature matrix (Fig. 2B). This indicates that the human brain has a specific population coding of mathematical operations independent of sensorimotor, language, or general executive load. Our results are also consistent with previous neuroimaging studies that reported predictability with both
sparse and latent features of music (Nakai et al., 2021) and cognitive tasks (Nakai and Nishimoto, 2020). Indeed, these two types of representations are at the core of how we model brain activity, regardless of the cognitive domains of interest.

The RSA revealed different similarity patterns between the sparse operator model and three types of latent features (ANN, word2vec, motion energy, Fig. 3). ANN features showed the largest correlations in the bilateral IPS, in line with previous studies on math cognition (Arsalidou and Taylor, 2011; Nieder, 2016). In contrast to ANN features, word2vec (i.e. language) features showed the largest correlation in the left precentral sulcus (slightly posterior to the inferior frontal gyrus) and left-lateralised correlation patterns, whereas motion energy (i.e. visual) features showed the largest correlation in the bilateral occipital cortex. These results are consistent with previous reports on language and visual neurosciences (Bradshaw et al., 2017; Güçlü and van Gerven, 2015; Nishimoto et al., 2011), suggesting that different latent feature distributions represent distinct aspects of mathematical operations. In particular, the word2vec features likely captured the semantic components of numbers and operator symbols, which is in line with a previous neuroimaging study showing the contribution of the operators' semantic component (Pyke et al., 2017).

The FBS analysis enabled the reconstruction of a brain representation of the sparse operator model based on the latent ANN model (Fig. 4). Interestingly, the ANN model outperformed the other two latent feature models across whole-cortical voxels. The difference between the RSA and FBS results might be due to analytic procedures. The RSA calculates correlations across different operators for each anatomical ROI, whereas the FBS calculates correlations separately for each operator in each cortical voxel. Therefore, the FBS can reveal the finer organisation of the distributed information of mathematical operations in the brain, and is a promising way to bridge the gap between neuroimaging based on sparse coding and population coding (Foldiak, 2003; Thorpe, 2012).

The increase in correlation between operator-weight maps and FBS maps may reflect a representational hierarchy of mathematical operations in the ANN model (Fig. 4F, S4). Previous studies revealed a correspondence of representational hierarchy between ANNs and BNNs in the visual (Güçlü and van Gerven, 2015; Horikawa and Kamitani, 2017; Yamins et al., 2014), auditory (Kell et al., 2018; Koumura et al., 2019) and language domains (Caucheteux and King, 2022; Hannagan et al., 2021). Our result is in line with previous work, in that deeper ANN layers represent latent features underlying abstract cognitive processes such as mathematical operations.

In contrast to previous studies on operator decoding (Haynes et al., 2007; Knops et al., 2009; Kutter et al., 2022; Pinheiro-Chagas et al., 2019), we successfully decoded novel operators (Fig. 5). Such generalisation was achieved using latent ANN features because the latent vectors were represented in a high-dimensional feature space common to both learned and unlearned tasks. Note that the current decoding analyses do not simply classify learned and unlearned tasks, they provide information on how much the decoded vectors are specific to several candidate tasks not included in model training (Fig. 5B, 5C). Our models also showed generalised decomposability across math expressions with different structures (i.e. single- and double-operator problems). Additional analyses using decoding models trained without target operators demonstrated that the decoding results were not a product of shared operators in math expressions (e.g. ‘AddMul’ and ‘AddDiv’ share ‘Add’). These results indicate that the ANN model represents the brain activity pattern of mathematical operations as distributions in a high-dimensional latent feature space.

It is worth noting some limitations in the current study. First, we did not compare the performance of different ANNs (e.g. (Andor et al., 2019; Lample and Charton, 2019; Wang et al., 2019) for math problem solving. However, the current study did not aim to determine the best fitting ANN model for the brain and TP-transformer is advantageous as an interpretation of its latent vectors has already been reported (Russin et al., 2021). The best brain-like ANN model of mathematical operations should be determined in future studies using concepts such as the brain-score (Schröpf et al., 2020), while our FBS approach can be an alternative way for this aim.

Second, some low-level information might not be captured by the current non-mathematical features (sensorimotor, language, executive load). In addition, non-mathematical features may cause different performance degradation between operator and ANN models because the latent ANN features contain more abstract representations than the operator features, the latter being more closely related to visual and semantic information. More controlled experimental designs are needed to rigorously assess the potential effects of non-mathematical features.

Third, although the number of subjects was determined based on previous studies using encoding/decoding (Horikawa et al., 2020; Huth et al., 2016; Nakai and Nishimoto, 2020; Popham et al., 2021), the relatively small number of subjects involved may affect the current findings. Indeed, two subjects out of eight showed inconsistent patterns of prediction map similarity (Fig. 2C, D). These subjects, however, also showed a significant correlation between FBS and operator-weight maps (Fig. 4E), indicating that the representational relationship between the sparse operator and latent ANN models is robust across individuals. Further research on a larger dataset may clarify individual differences in the distributed brain representations of mathematical operations. The current study is a first step toward ANN modelling of the brain representations of complex human cognitive abilities, such as mathematical thought, bridging the gap between symbolic and distributed representations found in human neuroimaging.

5. Ethics statement

Written informed consent was obtained from all subjects prior to their participation in the study; the study was approved by the ethics and safety committee of the National Institute of Information and Communications Technology in Osaka, Japan.

Declaration of Competing Interest

The authors declare no competing interests.

Credit authorship contribution statement

Tomoya Nakai: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft. Shinji Nishimoto: Supervision, Writing – review & editing, Funding acquisition.

Data Availability

The source data and analysis code used in the current study are available from Zenodo (https://doi.org/10.5281/zenodo.6614683).

Acknowledgements

We thank MEXT/JSPS KAKENHI (grant numbers JP17K13083 and JP18H05091 in #4903 (Evolving)udies for T.N., JP15H05311 and JP18H05522 for S.N.), JST CREST JPMJCR18A5, ERATO JPMJER1801 (for S.N.), and H2020 Marie Sklodowska-Curie Actions (grant number 101023033 for T.N.) for partial financial support of this study. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.119980.
References

Andor, D., He, L., Lee, K., Pitter, E., 2019. Giving BERT a Calculator: finding Operations and Arguments with Reading Comprehension. arXiv [cs.CL].

Araújo, M., Taylor, M., 2011. Meta-analyses of brain areas needed for numbers and calculations. NeuroImage 54, 2382–2393.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300.

Brenda, A.R., Thompson, P.A., Wilson, A.C., Bishop, D.V.M., Woodhead, Z.V.J., 2017. Measuring language lateralisation with different language tasks: a systematic review. PeerJ 5, e9299.

Cautelaux, C., King, J.-R., 2022. Brains and algorithms partially converge in natural language processing. Commun. Biol. 5, 134.

Folkad, P., 2003. Sparse coding in the primate cortex. The handbook of brain theory and neural networks.

Gao, J.S., Huth, A.G., Lescroart, M.D., Gallant, J.L., 2015. Pycorpus: an interactive surface visualizer for fMRI. Front. Neuroinform. 9, 23.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nasbate, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fandl, L., Doyle, W., Friedman, D., Dogan, P., Melloni, L., Reichert, R., Devote, S., Flinker, A., Hasenfratz, I., Levy, O., Hasidim, A., Brenner, M., Maer, T., Norman, K.A., Devinsky, O., Hasson, U., 2022. Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380.

Groen, I.L., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.L., 2018. Distinct contributions of functional and deep neural networks features to representational similarity of scenes in human brain and behavior. Elle 7, e25962.

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

Hannagan, T., Agrawal, A., Cohen, L., Dehaene, S., 2021. Emergence of a compositional neural code for written words: recycling of a convolutional neural network for reading. Proc. Natl. Acad. Sci. U. S. A. 118(4), doi:10.1073/pnas.2014779118.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17, 323–328.

Horikawa, T., Cowen, A.S., Kolter, D., Kamitani, Y., 2020. The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. iScience 23, 101060.

Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchically visual features. Nat. Commun. 8, 15037.

Huth, A.G., de Heer, W.A., Griffits, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuroreport 76, 1210–1224.

Ischebeck, A., Zamaran, L., Sündertopf, C., Koppelrätter, F., Benke, T., Felber, S., Delazer, M., 2006. How specifically do we learn? Imaging the learning of multiplication and subtraction. Neuroimage 30, 1365–1375.

Jain, S., Huth, A., 2018. Incorporating context into language encoding models for fMRI. Adv. Neural Inf. Process. Syst. 31.

Kay, K.N., David, E.V., Prenger, R.J., Hansen, K.A., Gallant, J.L., 2008a. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum. Brain Mapp. 29, 142–156.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008b. Identifying natural images from human brain activity. Nature 452, 58–62.

Kell, A.J.E., Yamins, D.I.K., Shook, E.N., Norman-Haiguen, S.V., McDermott, J.H., 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98, 630–644.e16.

Kietzmann, T.C., Spoerer, J.C., Sörensen, L.K.A., Gick, R.M., Haak, O., Kriegeskorte, N., 2019. Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. U. S. A. 116, 21854–21863.

Kim, G., Jiang, J., Baek, S., Song, M., Park, S.-B., 2021. Visual number sense in untrained deep neural networks. Sci. Adv. 7, eabd1277.

Knoops, A., Thirion, B., Hubbard, E.M., Michel, V., Dehaene, S., 2009. Recruitment of an area involved in eye movements during mental arithmetic. Science 324, 1583–1585.

Koide-Majima, N., Nakai, T., Nishimoto, S., 2020. Distinct dimensions of emotion in the human brain and their representation on the cortical surface. Neuroimage 222, 117258.

Kouwara, T., Terasaka, H., Fukurawa, S., 2019. Cascaded Tuning to Amplitude Modulation for Natural Sound Perception. J. Neurosci. 39, 5517–5533.

Kutter, E.F., Boström, J., Elger, C.E., Nieder, A., Mormann, F., 2022. Neuronal codes for arithmetical rule processing in the human brain. Curr. Biol. 32, 1275–1284.

Lample, G., Charlot, F., 2019. Deep Learning for Symbolic Mathematics. arXiv [cs.SCI].

Moffat, A., Turner, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. Advances in Natural Information Processing Systems.

Nakai, T., Koide-Majima, N., Nishimoto, S., 2021. Correspondence of categorical and feature-based representations of music in the human brain. Brain Behav 11, e01936.