Learning Structure from the Ground up—Hierarchical Representation Learning by Chunking

Shuchen Wu
Computational Principles of Intelligence Lab
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
shuchen.wu@tuebingen.mpg.de

Noémi Éltető
Department of Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
noemi.elteto@tuebingen.mpg.de

Ishita Dasgupta
Computational Cognitive Science Lab
Department of Psychology
Princeton University
dasgupta.ishita@gmail.com

Eric Schulz
Computational Principles of Intelligence Lab
Max Planck Institute for Biological Cybernetics
Tübingen, Germany
eric.schulz@tuebingen.mpg.de

Abstract

From learning to play the piano to speaking a new language, reusing and recombining previously acquired representations enables us to master complex skills and easily adapt to new environments. Inspired by the Gestalt principle of *grouping by proximity* and theories of chunking in cognitive science, we propose a hierarchical chunking model (HCM). HCM learns representations from non-i.i.d. sequential data from the ground up by first discovering the minimal atomic sequential units as chunks. As learning progresses, a hierarchy of chunk representations is acquired by chunking previously learned representations into more complex representations guided by sequential dependence. We provide learning guarantees on an idealized version of HCM, and demonstrate that HCM learns meaningful and interpretable representations in a human-like fashion. Our model can be extended to learn visual, temporal, and visual-temporal chunks. The interpretability of the learned chunks can be used to assess transfer or interference when the environment changes. Finally, in an fMRI dataset, we demonstrate that HCM learns interpretable chunks of functional coactivation regions and hierarchical modular and sub-modular structures supported by the neuroscientific literature. Taken together, our results show how cognitive science in general and theories of chunking in particular can inform novel and more interpretable approaches to representation learning.

1 Introduction

Sequential data in our everyday life is often hierarchically structured. From streaming this sequential sensory perceptual data, we can identify repeated patterns – and bootstrap these to recognize higher order patterns. In cognitive science, identifying repeated, invariant patterns from sequences in units is known as *chunking*. To get an intuition for chunking, try to read through the following sequence

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

https://doi.org/10.52202/068431-2660

of letters: “DFJKJKJKDFDFJKJKDFDF”. Upon reaching the end, if you were asked to repeat the letters from memory, you might recall fragments of the sequence such as “DF” or “JK”. By parsing the sequence of letters only once, you have already detected frequently occurring patterns and memorized them together as units, i.e. *chunks*. Chunking has been observed in a range of sensory and behavioral modalities including language learning [1, 2], action organization [3, 4] and visual perception of structures [5–7]. Chunking as a mechanism is a basis for humans to identify patterns as objects, assigning labels to them to facilitate memory compression [8, 9], sequence prediction [10, 11], communication [12, 13], and generalization[14]. Learning hierarchical representations of the world is a feature central to human intelligence.

Despite recent success, deep learning models, on the other hand, do not represent explicit hierarchies. Neural networks contain sub-symbolic, nested, non-linear structures whose prediction processes are hard to comprehend. This lack of interpretability raises concerns over their fairness, privacy, robustness and trust-worthiness [15–18] and manifests itself as a key shortcoming of these models [19, 20]. To address these shortcomings, researchers have urged to seek inspiration from cognitive science to construct models that resemble the hierarchical and interpretable representations as observed in human learners [21, 22]. We take a two-fold approach to this problem. First, instead of learning from iid data, we ask: what if the time series data that comprises streams of perception comes from a hierarchical structure? Under this assumption, what could be an algorithm that learns the embedded hierarchical structure? We take inspiration from models in cognitive science showing that people perceive structures based on the Gestalt principle of *grouping by proximity*, and formulate a generic hierarchical pattern discovery algorithm that enables the rational discovery of structures with embedded hierarchies. We refer to this model as the hierarchical chunking model (HCM).

HCM starts out learning a minimal set of units sufficient to explain the sequence and gradually combines these units into increasingly larger and more complex chunks, constructing interpretable hierarchical structures. We derive learning guarantees on an idealized generative model and demonstrate convergence on sequential data coming from this generative model. Thereby, Gestalt principles of grouping can be understood as a rational way of learning representations from sequences with an inherent hierarchical structure. We then show that HCM resembles more to human chunking in qualitative ways compared to a recurrent neural network and flexibly transfers components learned from one task to another. We extend HCM to the visual-temporal domain capable of learning visual-temporal parts and wholes from higher dimensional sequential data. Taking it one step further, we deploy HCM to learn from high-dimensional fMRI data, which exerts a hierarchical structure. We demonstrate HCM’s interpretable feature extraction ability to discover submodules of brain activations directly linkable to behavior supported by the literature.

2 Hierarchical Chunking Model

We define a chunk as a unit created by concatenating several atomic sequential units together. Taking the training sequence shown in Figure 1a as an example, the sequence is made up of discrete atomic units from an atomic alphabet set $\mathbb{A}_0$: in this case $\mathbb{A}_0 = \{0, 1, 2\}$. A chunk c is made up of a combination of one or more atomic units in $\mathbb{A}_0 \backslash \{0\}$. 0 denotes an empty observation in the sequence.

Intuitively, if a sequence contains inherent hierarchical structure, then there are patterns which span several sequential units sharing these internal structures, examples of such sequences are repeated melodies and sub-melodies in music. If the pattern occurs in the sequence, observations between sequential units within the pattern will be correlated. In this case, chunking patterns within a sequence as units simplifies perceptual processing in the sense that the sequence can be perceived one chunk after another, instead of one sequential unit at a time. Furthermore, the acquired “primary” chunks serve as building blocks to discover larger chunks that are embedded within the hierarchy of the sequential structure.

Formally, HCM acquires a belief set $\mathbb{B}$ of chunks, and uses chunks from the belief set to parse the sequence. HCM assumes that a sequence is generated from samples of independently occurring chunks with probability of $P_{\mathbb{B}}(c)$ evaluated on the belief set $\mathbb{B}$. The probability of observing a sequence of parsed chunks $c_1, c_2, ..., c_N$ can be denoted as $P(c_1, c_2, ..., c_N) = \prod_{c_i \in \mathbb{B}} P_{\mathbb{B}}(c_i)$. Chunks as perceiving units serve as independent factors that disentangle observations in the sequence.

https://doi.org/10.52202/068431-2660

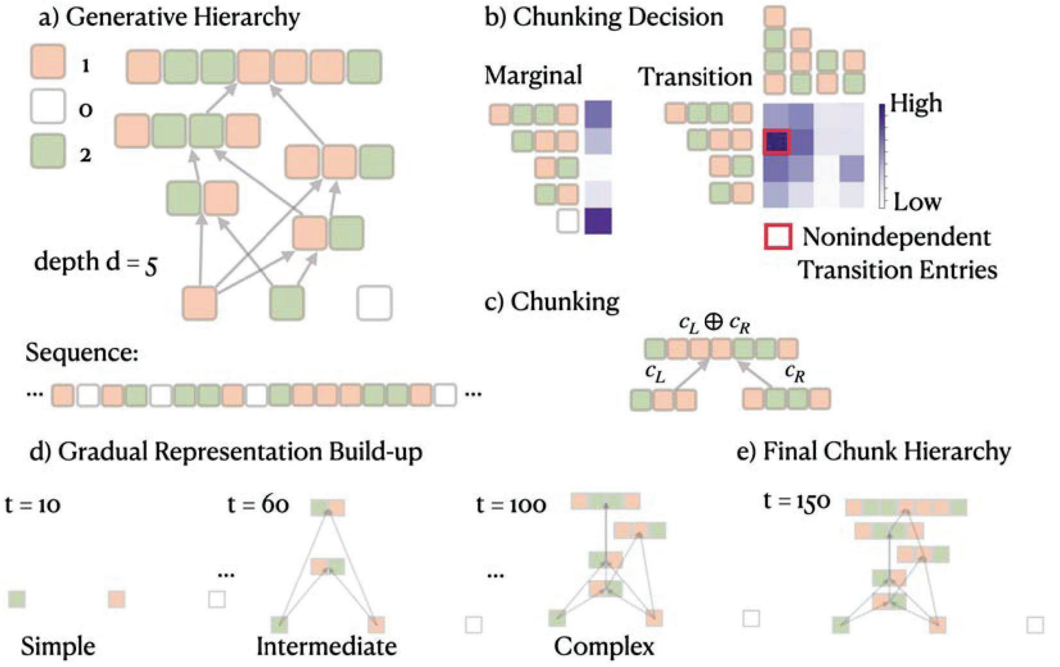

Figure 1: The Hierarchical Chunking Model. **a)** Example of a hierarchical model generating training sequences. **b)** Intermediate representation of learned marginal and transition matrices. The most frequent transition that violates the testing criterion is marked in red and can be turned into a new chunk. **c)** HCM combines the two chunks c_L and c_R to form a new chunk. **d)** As HCM observes longer sequences, it gradually learns a hierarchical representation of chunks. **e)** HCM arrives at the finally chunk hierarchy isomorphic to the generative hierarchy.

The training sequence is parsed by HCM in chunks. At every parsing step, the longest chunk in the belief set consistent with the upcoming sequence is chosen to explain the up-coming sequential observations. The end of the previous parse initiates the next parse.

Observing a hierarchically structured sequence as illustrated in Figure 1a, HCM gradually builds up a hierarchy of chunks starting from an empty belief set. It first identifies a set of atomic chunks to construct its initial belief set $\mathbb{B}$. Initially, these will be chunks of length one, yielding one-by-one processing of the primitive elements.

For one belief set $\mathbb{B}$, HCM keeps track of the marginal parsing frequency $\boldsymbol{M}(c_i)$ for each chunk c_i in $\mathbb{B}$, a vector with size $|\mathbb{B}|$ and the transition frequency $\boldsymbol{T}$ between chunk c_i followed by chunk c_j, as illustrated in Figure 1b. Entries in $\boldsymbol{M}$ and $\boldsymbol{T}$ are used to test the hypothesis that consecutive chunk parses have a correlated consecutive occurrence within the sequence via a χ^2-independence test. If two chunks c_L and c_R have a significant adjacency dependence based on their entries in $\boldsymbol{M}$ and $\boldsymbol{T}$, they are chunked together to become $c_L \oplus c_R$, which augments the belief set $\mathbb{B}$ by one. One example of chunk merging is shown in Figure 1c.

Independence Test We use a χ^2-test of independence to assess the correlation of consecutive occurrences of c_L followed by c_R. Let c_L be an indicator variable that is 1 when chunk c_l is parsed and 0 otherwise, similarly we formulate c_R as another indicator variable of parsing the chunk c_r. We evaluate the χ^2-value as a criterion to reject the null hypothesis that the consecutive observation of c_l followed by c_r is statistically independent:

$$\chi^2 = \sum_{c_L=\{0,1\}} \sum_{c_R=\{0,1\}} \frac{N(p(c_L, c_R) - p(c_L)p(c_R))^2}{p(c_L)p(c_R)}$$

$p(c_L, c_R)$ and $p(c_L)p(c_R)$ are evaluated from $\boldsymbol{M}$ an $\boldsymbol{T}$. The degree of freedom is 1. A χ^2-probability of less than 0.05 is the criterion to reject the null hypothesis (i.e. that c_l and c_r occur independently).

There are two versions of HCM. The Rational Chunk Learning HCM learns chunks in an idealized way which we use to study learning guarantees. The online version of HCM is an approximation

https://doi.org/10.52202/068431-2660

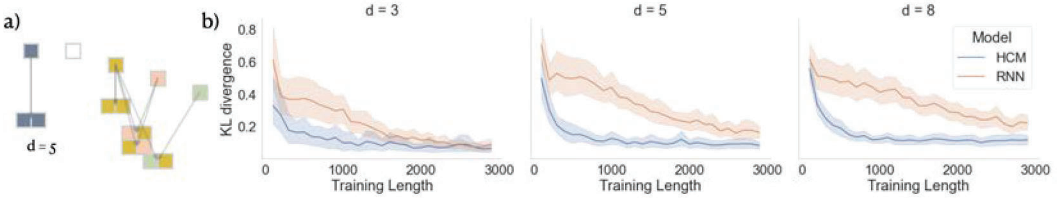

Figure 2: **a)** Example graph generated from the hierarchical generative model with a depth of $d = 5$. **b)** Learning performance of HCM and RNN with increasing training length and for increasing depths. Performance was averaged over 30 randomly-generated graphs.

to the rational HCM that can be adapted to different environments and processes sequences online. Pseudo-code for both algorithms can be found in the SI.

Rational Chunk Learning: HCM as an Ideal Observer HCM is initiated with an empty belief set and it first finds a minimally complete belief set after the first sequence parse. In each iteration, the entire sequence is parsed to evaluate $\boldsymbol{M}$ and $\boldsymbol{T}$, which are used to find consecutive chunk parses in the existing belief set that violate the independence testing criterion. From these dependent chunk pairs, the pair with the largest estimated joint probability is combined into a new chunk. The new chunk enlarges the belief set by one. The chunks in the new belief set are used to parse the sequence in the next iteration. This process repeats until all of the chunks in the belief set pass an independence halting criterion, which measures if all of the chunks in the belief set are currently independent, again assessed via a χ^2-test (see SI).

Online Chunk Learning The online chunk learning HCM approximates the ideal observer HCM by learning new chunks when the training sequence is processed on the go. To have a feature that encourages adaptation to new environmental statistics, entries in $\boldsymbol{M}$ and $\boldsymbol{T}$ can be subject to memory decay. We use the ideal observer model to demonstrate learning guarantees, but use the online model to learn representations in realistic and more complex set-ups.

2.1 HCM Learns Representations from the Ground Up

As HCM learns from a sequence, it starts with no representation and gradually builds up interpretable representations described by a chunk hierarchy graph $\hat{\mathcal{G}}$ with the vertex set being the chunks and edges pointing from constituents to composites. Shown in Figure 1d is the gradual build-up of one such graph as the model learns from a training sequence coming from the generative hierarchy in Figure 1a. At $t = 10$, HCM learns only the atomic chunks, at $t = 60$, HCM has already constructed two additional chunks; when $t = 100$, two more additional chunks are constructed. HCM arrives at the final chunk hierarchy at $t = 150$.

3 Generating Sequences with a Hierarchical Structure

We construct a generative model to study HCM's behavior formally and empirically. The generative model constructs random chunk hierarchies from which non-iid sequences are sampled. Such graph $\mathcal{G}_d$ contains vertex set $V_{\mathbb{A}_d}$ and edge set $E_{\mathbb{A}_d}$ to describe the relation between chunks and their constituents. One example is illustrated in Figure 1a. $\mathbb{A}_d$ is the set of chunks used to construct the sequence. The depth d specifies the number of chunks created in the generative process.

Starting with an initial set of atomic chunks $\mathbb{A}_0$, at the i-th iteration, two chunks c_L, c_R are randomly chosen from the current set of chunks $\mathbb{A}_i$ and are concatenated into a new chunk $c_L \oplus c_R$, augmenting $\mathbb{A}_i$ by one to $\mathbb{A}_{i+1}$. Meanwhile, an independent occurrence probability is assigned to each chunk under the constraint that the probability of occurrence for every new chunk c_i in the construction process evaluated on the support set $\mathbb{A}_i$ carries the largest probability mass.

Once a graph hierarchy is constructed, we construct non-iid observational sequences by consecutively sampling chunks from the hierarchy with their corresponding probability, under the constraint that no two chunks with a child chunk are sampled consecutively.

3.1 Learning Guarantee

Theorem: As the length of the sequence approaches infinity, HCM learns a hierarchical chunking graph $\hat{\mathcal{G}}$ isomorphic to the generative hierarchical graph $\mathcal{G}$.

Proof Sketch: We approach this proof by induction. Further details can be found in the SI. Base step: The first step of the rational chunking algorithm is to find the minimally complete atomic set of chunks to form its initial belief set. This procedure guarantees that $\hat{\mathcal{G}}_0 = \mathcal{G}_0$. Additionally, the probability mass of the learning model at step 0 and the generative model at step 0 is asymptotically the same as the sequence length approaches infinity. Induction hypothesis: Assume that the learned belief set $\mathbb{B}_i$ at step i contains the same chunks as the alphabet set $\mathbb{A}_i$ in the generative model, the chunk combination pair with the biggest evaluated joint occurrence probability violating the independence test is picked to be concatenated into a chunk to extend the belief set: this chunk is the same chunk node created by the hierarchical generative model. End step: The chunk learning process stops once the independence criterion is no longer violated. This is the case once the chunk learning algorithm has learned a belief set $\mathbb{B}_d = \mathbb{A}_d$.

3.2 Learning Convergence and Comparative Data-efficiency

To evaluate and show HCM's learning performance, we trained HCM to learn hierarchies of chunks from sequences generated by the hierarchical generative model. Shown in Figure 2 is HCM's learning performance as sequence length increases, averaged over 30 independently generated random graphs with the same depth d. One example of such graphs is shown in Figure 2a. Kullback-Leibler divergence was used to evaluate learning performance. To this end, learned hierarchies by HCM were used to generate sequences, which were then evaluated on the support set of the alphabet set in the generative model.

Figure 2b shows the KL-divergence between the learned and ground-truth distribution for increasing depths d of the generative graphs. For each depth, the KL-divergence was evaluated on 30 random generative models with sequence length increasing from 50 to 3000. Overall, the KL-divergence decreased as the training sequence length increased and converged with longer training sequences, showing a closer representation resemblance to the generative model.

A similar training and learning evaluation was conducted on a 3-layer Recurrent Neural Network (RNN) with 40 hidden units for comparison. As the length of the training sequence increased, the KL-divergence of RNN converged at a slower rate than HCM. This competitive advantage in data efficiency became more pronounced with increasing depth of the generative hierarchy.

4 HCM Resembles Human Chunk Learning

Here we compare the chunk learning behavior of HCM to the learning characteristics of humans. To that end, we used data collected from a sequence learning study by [23] with 47 participants under the license CC-BY 4.0. As shown in Figure 3a, the training sequence comprised chunks ABC and D, independently occurring with equal probability. The study assessed how participants built up chunk knowledge gradually. Participants' reaction times reflected that, after enough training, they were anticipating several upcoming sequence elements, suggesting that they have acquired longer chunks (Figure 3b, left) [23].

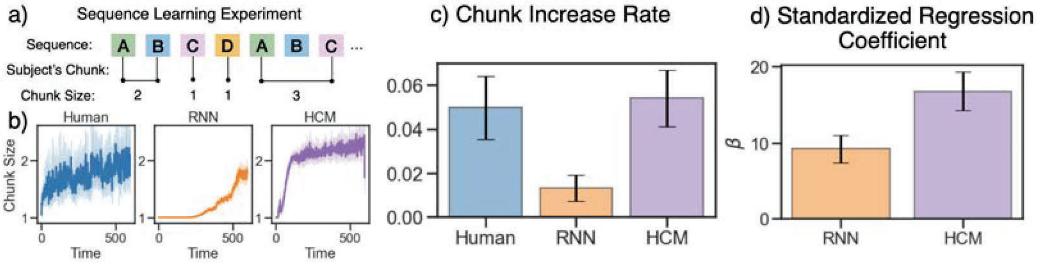

Figure 3: **a)** A sequence learning experiment with chunks ABC and D. **b)** Chunk size increase of human participants, RNN, and HCM during training. **c)** Average chunk increase rate during training. **d)** Regression coefficient of RNN and HCM's confidence estimates on human reaction time data.

In a similar vein, we measured online chunk size increase of HCM using the same method as in [23] and, for comparison, RNNs (see SI for further comparisons to other algorithms). HCM, similarly to humans, started learning longer chunks early in the sequence. By contrast, RNN did not start to chunk until after step 300, and when it started to learn chunks, the increase rate of the predictive horizon was not as steep as HCM's. Evaluating the average rate of chunk growth also showed that HCM builds up chunks as learning progresses was more similar to participants' than the RNN's (Figure 3c). The negative log-probabilities of sequence elements generated by the HCM and RNN were both significantly related to human reaction times (that reflect the certainty of their internal predictions [24]). Yet, the relationship was substantially stronger (Figure 3d) between HCM ($\beta = 16.74$, $p \leq 0.001$, $\tau = 0.165$, $BIC = 313586.5$) and human participants compared to that of the RNN ($\beta = 9.24$, $p \leq 0.001$, $\tau = 0.085$, $BIC = 314236.4$). These results suggest that HCM resembles human chunk learning more strongly than RNNs and can therefore be seen as the cognitively more plausible approach to hierarchical representation learning.

5 HCM Permits Transfer Between Environments

One characteristic of human learning is that previous learning experience facilitates and sometimes interferes with acquiring a new skill [25]. Having an interpretable representation can inform us about positive or negative transfer a priori.

An HCM has learned a chunking graph in Figure 4a from an environment. When it switches to another environment with a generative model overlapping with its previously acquired representation, it can reuse the learned subgraphs of chunks marked in gray as in Figure 4b and learns faster than a naive HCM in Figure 4c. Vice versa, transfer can be detrimental when there is no or little overlap between the learned chunking graph and the generative model of the new environment. For the same HCM as in Figure 4a, transferring to an environment with a chunking graph in Figure 4d implies learning the shaded chunks anew, in addition to running the risk of being misled by the previous representations. As a result, the early performance of the pre-trained HCM suffers more from an interfering environment than a naive model in Figure 4e. Interpretability of HCM's representations enables the assessment of facilitation or interference when the environment changes.

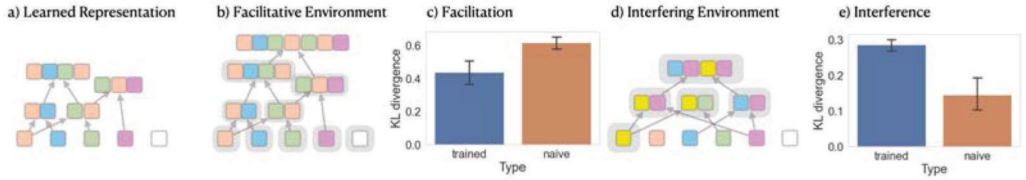

Figure 4: **a)** Example of a representation learned by an HCM. **b)** A facilitative environment with a generative model overlapping in its structure with the test environment. Gray shadows mark the chunks that can be directly transferred. **c)** Average performance over the first 500 trials after the environment switches. **d)** Interfering environment. Gray shadows marks chunks that need to be acquired anew. **e)** Average performance over the first 500 trials after the environment switches.

6 Generalizing to Visual Temporal Chunks via the Principle of Proximal Grouping

Humans excel at finding structures in hierarchical visual objects and grouped movements. The Gestalt principle of *grouping by proximity* suggests that we tend to group objects that are close to one another into a cohesive unit [26, 27]. This principle has been suggested to play a key role in human perceptual grouping [28], benefits working memory [29] and reduces visual complexity [30]. Indeed, in humans and other animals, learning of adjacent relationships prevails over non-adjacent ones [31]. Therefore, the adjacent dependency structure can be expanded to chunking in visual temporal domains [32]. To emulate this ability of chunking via proximal grouping, we extend HCM to learn visual temporal chunks.

Visual temporal chunks subsume temporal length and varying visual slices in each temporal slice (Figure 5a). One can imagine a visual temporal chunk as having a 3D shape — the first two

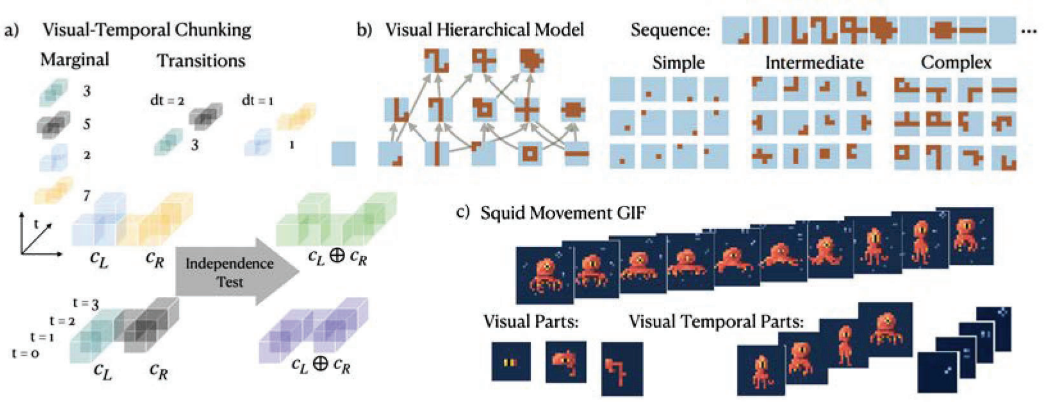

Figure 5: **a)** HCM learns visual-temporal chunks by extending the transition matrix to take account of time differences. **b)** Left: A visual hierarchical model where complex images are composed of simpler images. Right: Initial, intermediate, and complex chunks learned by HCM trained on sequences of images sampled from the visual hierarchical model.**c)** Top: A GIF of a moving visual used as a sequence to train HCM. Bottom: Examples of temporal and visual chunks learned by HCM.

dimensions are the visual part of the chunk, and the object's length is the temporal part, made of stacked visual-temporal pixels. Within each temporal slice are the visual features identified by the chunk. As the model iterates through data across its temporal slice, the chunk that attains the biggest visual temporal volume explains part of the observational sequence. Multiple visual temporal chunks can occur simultaneously. Starting at the visual temporal time point marked by the previous chunk, chunks are identified and stored in $\boldsymbol{M}$. The transition matrix $\boldsymbol{T}$ is modified to account for the temporal lag difference between adjacent chunk pairs within a proximity parameter and records the frequency that one chunk transitions into another for each time lag. Whenever a pair of adjacent chunks are identified, an independence hypothesis test evaluates whether the adjacent observation are correlated. Chunks that violate the hypothesis test are combined to parse future sequences.

Learning Part-Whole Relationship Between Visual Components We show HCM learned chunks in the visual domain from a sequence of independently sampled images. Figure 5b left shows a hierarchical generative model in the pixel-wise image domain. A set of elementary visual units in the lowest hierarchy level combines into intermediate and more complex visual units higher up in the hierarchy. All of the constructed elements in the hierarchy occurred independently according to a probability drawn from a Dirichlet flat distribution. Images in the hierarchy were independently sampled from the generative distribution to become the training sequence. In Figure 5b, right we show the chunk representations learned by HCM at different stages. Initially, HCM acquires the individual pixels as chunks to explain the observations. As HCM proceeds with learning, it discovers visual correlations among the pixels and constructs increasingly complex visual parts.

Learning Visual-Temporal Movement Hierarchies Instead of seeing one image after another sampled from an independent, identically distributed distribution, real-world experiences contain correlations in both the visual and temporal dimensions. From observing object movements across space and time, the visual system learns structures from correlated visual and temporal observations, decomposes motion structure, and groups moving objects together as a whole [33]. To emulate this type of environment, an animated GIF of a squid swimming in the sea (Figure 5c) was used as a visual-temporal sequence to train HCM. As learning advances, HCM learns chunks spanning both the visual and temporal domains. There are visual-temporal chunks that mark the movements of a tentacle and the rising-up motion of a bubble. Additionally, visual chunks resemble a part of the visual's eye and face. The meaningful chunks in the visual-temporal domain suggest the grouping principle enables the plausible learning of movement sequences and aids the perception of objects as wholes and their corresponding parts.

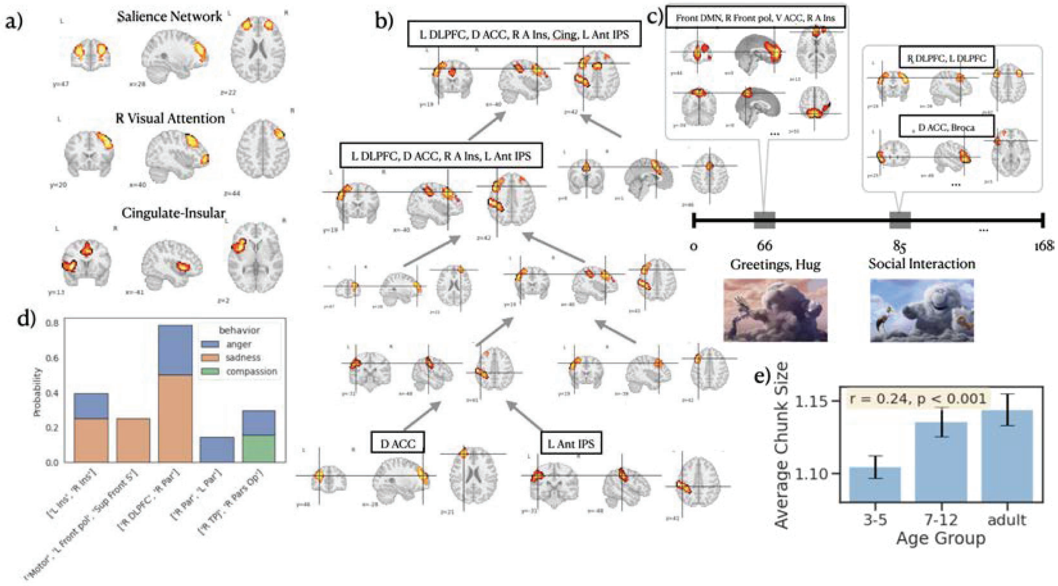

Figure 6: Application avenues of HCM on fMRI data. **a)** Example chunks of brain functional activation regions. **b)** HCM learns hierarchical functional network with bigger chunks emerging from its constituents. **c)** Chunk activation patterns responding to scene content **d)** Distinct response of retrieved chunks to tagged scenes. **e)** Average chunk size across age groups.

7 Learning Hierarchies of Brain Activation from Resting-state fMRI data

HCM learns hierarchies from structured sequential data. As brain activation has been suggested to be hierarchically structured [34], we demonstrate HCM's usefulness to learn structures in biological neural networks activating in response to complex stimuli by running HCM on a resting-state fMRI data set.

We used a developmental data set provided by the `nilearn` package with BSD License [35] and originally collected by [36] with its corresponding IRB approval. This data set contains the resting-state BOLD activity of 155 participants ranging from age 3 to 40, while watching the silent movie "Partly Cloudy". BOLD signal was extracted from functional brain regions defined by the MSDL Atlas [37], with confounds excluded and transformed into a rounded, normalized time series.

HCM's Chunks Reflect Structural, Functional and Anatomical Connectivity Figure 6 shows three typical examples of learned chunks for a randomly-chosen participant. The labels of functional regions come from the MSDL atlas [37]. The first example is the co-activation of D ACC and R A Ins. These two regions have been observed to co-activate in the presence of emotions, pain, and humor. They have been suggested to be a key hub of the salience network [38–41]. The second example chunk contains the activation of R DLPFC and R Front Pole. These regions belong to the visual attention network and are known to be anatomically connected [42–44]. A final example is the chunk of L Ins and Cing, which are also known to be anatomically and functionally connected [45]. Thus, the chunks discovered by HCM correspond to empirically-verified patterns of functional activity.

HCM's Chunks Recover Hierarchical Activation Patterns In fMRI data, hierarchies of chunk activation constructed by HCM reflect networks of functional regions. On the top of the hierarchy, the largest chunk contained L DLPFC, D ACC, R A Ins, Cing, and L Ant IPS (Figure 6b). Those regions are known to co-activate during cognitive tasks that demand attention, working memory, and control [41]. Chunks in the intermediate levels of the hierarchy reflect sub-networks of functional connectivity. Sub-chunks such as D ACC, R A Ins, L Ant IPS, and L DLPFC have been suggested to conjointly activate in cognitive effort-related activities [41]. Atomic chunks in the hierarchy such as D ACC and L Ant IPS activate individually sometimes without their parent chunks. Indeed, they have distinct functional signatures for affect processing [46, 47], and visual attention control [48]. Upon exposure to a time series in fMRI data, HCM constructs chunks from their constituents and arrives at a hierarchy of chunk relations, indicating nested network structures in the human brain.

HCM's Chunks can be Matched with Stimulus Onsets The retrieved chunks by HCM can be tagged with critical stimulus onsets. We tagged 19 critical moments involving social and emotional content in the movie. Figure 6c shows one example chunk activation upon stimulus onsets. Frontal DMN, right frontal pole, ventral anterior cingulate cortex, and right anterior insula activate together as a recurring unit after participants witness a scene with characters greeting and hugging each other. These regions have been suggested to be involved in social and cognitive processing [49]. Another example is the activation of areas known to be involved in emotion and language processing: D ACC and Broca [50], during a scene containing social interactions. In the meantime, the left and right prefrontal cortex, involved in theory-of-mind [36], also lights up.

We categorized the tagged moments into 3 groups of different emotional load: sadness, anger, and compassion. We then looked at the activation probability of retrieved chunks within the 6 seconds after watching those tagged scenes. Figure 6d shows a list of such chunks from one participant with their activation probability for each emotional category. For example, the left and right insula, known to be involved in affective processing [51], have a 0.4 activation probability after witnessing scenes of sadness or anger, but no activation after witnessing scenes of compassion. The same holds for R DLPFC and R Par that have been documented to activate in response to emotional conflict [52]. On the other hand, regions such as R TPJ and R pars opercularis that are involved in emotional reactions and theory of mind processes [53], activate in response to a scene of compassion or anger, but not to a scene of sadness. Thus, chunks of active brain regions can be related to complex stimuli, and regions activate selectively in response to one or more categories of emotional stimuli, but not others.

HCM's Average Chunk size Correlates with Participants' Age HCM can also be used to perform meaningful analyses at the population level. Specifically, we find that HCM's returned average chunk size per participant correlates significantly with age (Figure 6e). The older the participants are, the longer are the chunks found in their data ($r = 0.23$, $p \leq 0.001$). This discovery is in line with findings in the original study, which showed an increase in modularization of ToM and pain circuits across development [36].

To summarize, we applied HCM to learn chunks from a developmental fMRI data set. HCM enabled the discovery of spatially and temporally correlated activation chunks that are theoretically and empirically meaningful. The resulting chunks can be linked to complex stimuli and offer directly interpretable insights into the structure and function of brain activity.

8 Related Work

HCM extends upon decades of previous cognitive science and psychology research on chunking. In cognitive science, process models such as PARSER and competitive chunking were demonstrated to generate qualitatively similar chunks as in human sequence learning [54, 55]. HCM is a rational algorithm that learns the underlying chunks when the sequence is generated from a hierarchical chunking graph. Therefore, the chunking criterion is no longer a heuristic but a rational learning strategy that enables hierarchical structural discovery. On top of inheriting the merits of its predecessors, HCM generalizes the chunk learning principle to higher dimensional sequential domains such as visual-temporal sequential data.

HCM relates to several other lines of research. One is program induction. In program induction, explicit representations are acquired by searching for programmatic structures that best explain observational samples [22], and consolidating these offline with library learning [56]. However, domain expert knowledge is needed to specify the primitive programs; the relations and composition rules must adapt to the task settings and sensitively influences the quality of retrieved representations. Other approaches to structure learning include unsupervised parsing [57], which learns a stochastic and-or graph from sequential data. HCM is distinct in adapting its representation granularity to discover bigger chunks from data without pre-specifying the structure.

Another category of models to learn from sequential data are traditional sequence learning models including Hidden Markov Models (HMM), n-gram models and their variants to capture multi-scale sequential structure such as Hidden semi-Markov model [58] and hierarchical HMM [59]. The parameters of these models proliferate exponentially as a function of chunk length, implying memory inefficiency. Additionally, these models demand a structure specification before fitting parameters to the data. They also lack the adaptive recombination and reuse of pre-existing components. The

https://doi.org/10.52202/068431-2660

same issue is with neural network approaches to extract chunks from sequences [60–62]. Apart from lacking in interpretability, these models do not leverage the concatenation process observed in humans or reusing previously learned representations to construct new representations.

The principle of iterative merging of chunks has been used in compression algorithms, such as tokenizing methods in NLP, which were developed to optimize sequential data compression. Tokenizing methods such as Byte-Pair Encoding [63] iteratively merges the most frequent pairs of chunks to build a vocabulary of a text corpus. This objective is easy to compute but gives rise to ambiguous parses of the text (e.g. [AB, C] / [A, BC]). To minimize parse ambiguity, WordPiece [64] merges chunks that increase the likelihood of the corpus the most. However, the objective of WordPiece is expensive to compute. HCM circumvents the problem of computing the global sequence likelihood by instead maximizing the local chunk continuation likelihood. The computational efficiency of HCM makes it a plausible cognitive model of chunking as well as a promising method for NLP.

Probabilistic context-free grammars (PCFGs) are related to HCM in that they use trees as a representational form of sequences. The parse trees of PCFGs denote production rules, such as S $\rightarrow$ NP + VP. These production rules define how abstract syntactic units (non-terminals), such as a noun phrase and a verb phrase, are instantiated into a concrete string of words (terminals) to compose a sentence, such as 'we wrote the paper'. In comparison, the generative tree of HCM denotes statistical relationships among concrete chunks, such as 'we'-'wrote'-'the paper'. Extending HCM to represent abstraction is an exciting future direction, on which avenue the comparison to PCFGs will be instructive.

9 Discussion

Our work has its limitations. Currently, we fix the memory decay and the deletion threshold parameters to a priori plausible values. In future work, these parameters could be adapted online based on environment volatility. Another limitation is its scalability: at the moment, HCM learns representations from semi-high dimensional sequential data (i.e., currently between 1 to 625 dimensions). We are actively looking into generalizing this algorithm to higher dimensional data domains by combining it with existing neural network approaches or computer vision algorithms such as coherent point drift [65] or normalized cuts [66] to allow for the learning of ambiguous and high dimensional chunk exemplars. It is also possible to combine HCM with the compressed representation, such as the hidden activity of an auto-encoder to process and learn the structure from downstream representation. In this work, HCM learns one type of hierarchy of compound representations. However, we can show that HCM can be generalized to not only learn simple chunks but also chunks in projected spaces and thereby generalize between two chunks that contain the same motif (for example, "12221212" and "34443434"; see SI for detailed results). In the future, it might be worthwhile to further combine our approach with others amongst a taxonomy of representational hierarchies.

HCM also opens up other application directions. One direction is integrating HCM with deep neural network approaches as an interface between human understanding and distributed computation. Learning hierarchies of coherent activations from intermediate hidden units has the potential to reveal neural networks' underlying computation structure. Furthermore, it is also possible to equip HCM with additional top-down encoded representations, for example, by pre-training on other sequences or by adjusting the chunks by hand before the training starts. Another direction in neuroscience or behavioral research is to learn chunks of tagged animal movements that enable insights into the emergence of behavioral structure [67]. Finally, finding patterns that form as a cognitive unit is a vital task for infants to learn about the structures of the world and resembles the process of formulating a scientific theory from observation [68]. HCM can function as one means to come up with world models by observation, ready for experimental interventions or active learning to delineate the causal structure within [69].

10 Conclusion

We have proposed a hierarchical chunking model (HCM) that learns chunks from non-iid sequential data with a hierarchical structure. HCM starts out learning an atomic set of chunks to explain the sequence and gradually combines them into increasingly larger and more complex chunks. The output of the model is a dynamical graph that is a trace of the evolving representation. The resulting representations are easy to interpret, and flexibly reusable.

https://doi.org/10.52202/068431-2660

References

[1] Pierre Perruchet, Bénédicte Poulin-Charronnat, Barbara Tillmann, and Ronald Peereman. New evidence for chunk-based models in word segmentation. *Acta psychologica*, 149:1–8, 2014.

[2] Stewart M McCauley and Morten H Christiansen. Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, 9(3):637–652, 2017.

[3] Virginia B. Penhune and Christopher J. Steele. Parallel contributions of cerebellar, striatal and M1 mechanisms to motor sequence learning, 2012. ISSN 01664328.

[4] David A. Rosenbaum, Sandra B. Kenny, and Marcia A. Derr. Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 1983. ISSN 00961523. doi: 10.1037/0096-1523.9.1.86.

[5] Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery*. *Cognitive Science*, 3(3):231–250, 1979. doi: https://doi.org/10.1207/s15516709cog0303_3. URL `https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0303_3`.

[6] Timothy F. Brady, Talia Konkle, and George A. Alvarez. Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations. *Journal of Experimental Psychology: General*, 138(4), 2009. ISSN 00963445. doi: 10.1037/a0016797.

[7] Dennis E. Egan and Barry J. Schwartz. Chunking in recall of symbolic drawings. *Memory & Cognition*, 7(2), 1979. ISSN 0090502X. doi: 10.3758/BF03197595.

[8] Fernand Gobet, Peter C.R. Lane, Steve Croker, Peter C.H. Cheng, Gary Jones, Iain Oliver, and Julian M. Pine. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 2001. ISSN 13646613. doi: 10.1016/S1364-6613(00)01662-4.

[9] George A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 1956. ISSN 0033295X. doi: 10.1037/h0043158.

[10] Iring Koch and Joachim Hoffmann. Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychological Research*, 63(1), 2000. ISSN 14302772. doi: 10.1007/PL00008165.

[11] Diana Mussgens and Fredrik Ullén. Transfer in motor sequence learning: Effects of practice schedule and sequence context. *Frontiers in Human Neuroscience*, 11 2015. doi: 10.3389/fnhum.2015.00642.

[12] Ludwig Josef Johann Wittgenstein. *Philosophical Investigations*. New York, NY, USA: Wiley-Blackwell, 1953.

[13] Eric Schulz, Francisco Quiroga, and Samuel J Gershman. Communicating compositional patterns. *Open Mind*, 4:25–39, 2020.

[14] Eric Schulz, Joshua B. Tenenbaum, David Duvenaud, Maarten Speekenbrink, and Samuel J. Gershman. Compositional inductive biases in function learning. *Cognitive Psychology*, 2017. ISSN 00100285. doi: 10.1016/j.cogpsych.2017.11.002.

[15] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1:1–10, 10 2017.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[18] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. doi: 10.1109/TETCI.2021.3100641.

[19] B. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*, 2018.

[20] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. doi: 10.1016/0010-0277(88)90031-5.

[21] François Chollet. On the measure of intelligence, 2019. URL `https://arxiv.org/abs/1911.01547`.

[22] B. Lake, R. Salakhutdinov, and J. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332 – 1338, 2015.

[23] Shuchen Wu, Noémi Éltető, Ishita Dasgupta, and Eric Schulz. E pluribus unum but how? chunking as a rational solution to the speed-accuracy trade-off, Feb 2022. URL `psyarxiv.com/sjh27`.

[24] Claude Bonnet, Jordi Fauquet, and Santiago Ferrer. Reaction times as a measure of uncertainty. *Psicothema*, 20:43–8, 03 2008.

[25] Scott Jarvis and Aneta Pavlenko. Crosslinguistic influence in language and cognition. *Crosslinguistic Influence in Language and Cognition*, pages 1–287, 01 2007. doi: 10.4324/9780203935927.

[26] Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, and Cees Van Leeuwen. A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218, 2012.

[27] W. Metzger. *Laws of seeing*. MIT Press, 2006.

[28] Brian J Compton and Gordon D Logan. Evaluating a computational model of perceptual grouping by proximity. *Perception & Psychophysics*, 53(4):403–421, 1993.

[29] Dwight J Peterson and Marian E Berryhill. The gestalt principle of similarity benefits visual working memory. *Psychonomic bulletin & review*, 20(6):1282–1289, 2013.

[30] Don C Donderi. Visual complexity: a review. *Psychological bulletin*, 132(1):73, 2006.

[31] Raphaëlle Malassis, Arnaud Rey, and Joël Fagot. Non-adjacent dependencies processing in human and non-human primates. *Cognitive Science*, 42(5):1677–1699, 2018.

[32] Vicky Froyen, Jacob Feldman, and Manish Singh. Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, 122(4):575, 2015.

[33] Johannes Bill, Hrag Pailian, Samuel J. Gershman, and Jan Drugowitsch. Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences*, 117(39):24581–24589, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2008961117. URL `https://www.pnas.org/content/117/39/24581`.

[34] Pedro Alves, Chris Foulon, Vyacheslav Karolis, Danilo Bzdok, Daniel Margulies, Emmanuelle Volle, and Michel Thiebaut de Schotten. An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings. *Communications Biology*, 2:1–14, 10 2019. doi: 10.1038/s42003-019-0611-3.

[35] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00014. URL `https://www.frontiersin.org/article/10.3389/fninf.2014.00014`.

[36] Hilary Richardson, Grace Lisandrelli, Alexa Riobueno-Naylor, and Rebecca Saxe. Development of the social brain from age three to twelve years. *Nat Commun*, 1027(9), 2018. doi: https://doi.org/10.1038/s41467-018-03399-2.

[37] Gael Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Proceedings of the 22nd International Conference on Information Processing in Medical Imaging*, IPMI'11, page 562–573, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642220913.

[38] William W. Seeley. The salience network: A neural system for perceiving and responding to homeostatic demands. *Journal of Neuroscience*, 39(50):9878–9882, 2019. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1138-17.2019. URL `https://www.jneurosci.org/content/39/50/9878`.

[39] Critchley HD. Neural mechanisms of autonomic, affective, and cognitive integration. *J Comp Neurol.*, 493(1), 2005. doi: 10.1002/cne.20749.

[40] Nick Medford and Hugo Critchley. Conjoint activity of anterior insular and anterior cingulate cortex: Awareness and response. *Brain structure & function*, 214:535–49, 06 2010. doi: 10.1007/s00429-010-0265-x.

[41] Bart Aben, Cristian Buc Calderon, Eva Van den Bussche, and Tom Verguts. Cognitive effort modulates connectivity between dorsal anterior cingulate cortex and task-relevant cortical areas. *Journal of Neuroscience*, 40(19):3838–3848, 2020. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2948-19.2020. URL `https://www.jneurosci.org/content/40/19/3838`.

[42] Kathleen J Burman, David H Reser, Hsin-Hao Yu, and Marcello G P Rosa. Cortical input to the frontal pole of the marmoset monkey. *Cerebral cortex (New York, N.Y. : 1991)*, 8(21), 2011. doi: https://doi.org/10.1093/cercor/bhq239.

[43] Huaigui Liu, Wen Qin, Wei Li, Lingzhong Fan, Jiaojian Wang, Tianzi Jiang, and Chunshui Yu. Connectivity-based parcellation of the human frontal pole with diffusion tensor imaging. *The Journal of neuroscience*, 16(33), 2013. doi: https://doi.org/10.1523/JNEUROSCI.4882-12.2013.

[44] Michael Petrides and Deepak N. Pandya. Efferent association pathways from the rostral prefrontal cortex in the macaque monkey. *The Journal of neuroscience*, 27(43), 2007. doi: https://doi.org/10.1523/JNEUROSCI.2419-07.2007.

[45] Keri S. Taylor, David A. Seminowicz, and Karen D. Davis. Two systems of resting state connectivity between the insula and cingulate cortex. *Human Brain Mapping*, 30(9):2731–2745, 2009. doi: https://doi.org/10.1002/hbm.20705. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20705`.

[46] Francis Stevens, R.A. Hurley, and Katherine Taber. Anterior cingulate cortex: Unique role in cognition and emotion. *Journal of Neuropsychiatry and Clinical Neurosciences*, 23:121–125, 01 2011. doi: 10.1176/jnp.23.2.jnp121.

[47] Jue Wang, Ning Yang, Wei Liao, Han Zhang, Chao-Gan Yan, Yu-Feng Zang, and Xi-Nian Zuo. Dorsal anterior cingulate cortex in typically developing children: Laterality analysis. *Developmental Cognitive Neuroscience*, 15:117–129, 2015. ISSN 1878-9293. doi: https://doi.org/10.1016/j.dcn.2015.10.002. URL `https://www.sciencedirect.com/science/article/pii/S1878929315000924`.

[48] Sarah Vinette and Signe Bray. Variation in functional connectivity along anterior-to-posterior intraparietal sulcus, and relationship with age across late childhood and adolescence. *Developmental Cognitive Neuroscience*, 31, 04 2015. doi: 10.1016/j.dcn.2015.04.004.

[49] Marisa Loitfelder, Stephan CJ Huijbregts, Ilya Milos Veer, Hanna S Swaab, Mark A Van Buchem, Reinhold Schmidt, and Serge A Rombouts. Functional connectivity changes and executive and social problems in neurofibromatosis type i. *Brain connectivity*, 5(5):312–320, 2015.

[50] A Craig. How do you feel—now? the anterior insula and human awareness. *Nature reviews. Neuroscience*, 10:59–70, 02 2009. doi: 10.1038/nrn2555.

[51] Lucina Uddin, Jason Nomi, Benjamin Hébert-Seropian, Jimmy Ghaziri, and Olivier Boucher. Structure and function of the human insula. *Journal of Clinical Neurophysiology*, 34:300–306, 07 2017. doi: 10.1097/WNP.0000000000000377.

[52] Francesca De Luca, Manuel Petrucci, Bianca Monachesi, Michal Lavidor, and Anna Pecchinenda. Asymmetric contributions of the fronto-parietal network to emotional conflict in the word–face interference task. *Symmetry*, 12(10):1701, 2020.

[53] Richard P. Bagozzi, Willem J. M. I. Verbeke, Roeland C. Dietvorst, Frank D. Belschak, Wouter E. van den Berg, and Wim J. R. Rietdijk. Theory of mind and empathic explanations of machiavellianism: A neuroscience perspective. *Journal of Management*, 39(7):1760–1798, 2013. doi: 10.1177/0149206312471393. URL https://doi.org/10.1177/0149206312471393.

[54] Pierre Perruchet and Annie Vinter. Parser: A model for word segmentation. *Journal of Memory and Language*, 39(2):246 – 263, 1998. ISSN 0749-596X. doi: https://doi.org/10.1006/jmla.1998.2576. URL http://www.sciencedirect.com/science/article/pii/S0749596X98925761.

[55] Emile Servan-Schreiber and John Anderson. Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:592–608, 07 1990. doi: 10.1037/0278-7393.16.4.592.

[56] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*, 2020.

[57] Kewei Tu, Maria Pavlovskaia, and Song-Chun Zhu. Unsupervised structure learning of stochastic and-or grammars. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/24681928425f5a9133504de568f5f6df-Paper.pdf.

[58] Shun-Zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2009.11.011. URL https://www.sciencedirect.com/science/article/pii/S0004370209001416. Special Review Issue.

[59] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

[60] Katrin Ortmann. Chunking historical german. In *NODALIDA*, 2021.

[61] Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. Neural models for sequence chunking. In *AAAI*, 2017.

[62] Sun Si, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, and Jie Bao. Joint keyphrase chunking and salience ranking with bert, 04 2020.

[63] Philip Gage. A new algorithm for data compression. http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM. Accessed: 2022-07-29.

[64] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. 09 2016.

[65] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. doi: 10.1109/TPAMI.2010.46.

https://doi.org/10.52202/068431-2660

[66] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, aug 2000. ISSN 0162-8828. doi: 10.1109/34.868688. URL `https://doi.org/10.1109/34.868688`.

[67] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[68] Alison Gopnik. The scientist as child. *Philosophy of Science*, 63(4):485–514, 1996. doi: 10.1086/289970.

[69] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] As this work introduces an algorithmic approach, it is not clear to us with the potential negative societal impact it can induce.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] A proof sketch is described in the main paper, while the full proof can be found in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They will be included in the supplementary material and publicly available.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details on the experiments will be included in the supplementary.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Results obtained run without special computing resources.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] From our awareness, no personally identifiable information is present in any of the assets.
5. If you used crowdsourcing or conducted research with human participants...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We used existing data from experiments conducted in previous work. This information can be found in the original articles.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] This question is not applicable to this project.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Since no human experiment is conducted for this project and we used experimental data from other published and publicly available work, this question is not applicable to us here and can be found in the original articles.

https://doi.org/10.52202/068431-2660