Towards Understanding and Developing Open-Ended Intelligences for Infinite Worlds

Omid Madani

San Carlos, CA, USA,
omidmadani@yahoo.com,
homepage: http:www.omadani.net/PGs_page.html

Abstract. We describe the state of our work on Prediction Games for unsupervised cumulative learning of structured perceptual concepts. Here, concepts predict one another and are built from each other. Improving at prediction drives the learning, and co-occurrences drive concept construction. In each episode, through the process of interpretation, the system determines which of its many concepts are useful, *i.e.* form a coherent account of (low-level) buffer contents. By practicing many interpretations, prediction weights are continually updated, and from time to time new concepts are generated, leading to improved predictions and more coherent accounting. Over the past few years, our approach has become more probabilistic and information-theoretic. We report on our improved results in recovering good split (*e.g.* word) boundaries, when starting at character or lower levels. We describe our current understanding of the challenges, including internal and external non-stationarity, and incorporating new concepts, as well as potential applications.

Keywords: Learning and Development, Open-Ended Learning, Cumulative (Hierarchical) Learning, Self-Supervised Learning, Perceptual Concepts, Interpretation, Constructivism, Semiotics, Learning for Perception

> *"... any visual input is consistent with an unlimited number of interpretations. The challenging question is how the perceptual system functions such that we normally are unaware of any ambiguity. Our experience is simply that of seeing things the way they are... our conceptual systems comes prepared with expectations.. These sorts of expectations or "constraints" occur in all facets of cognition."*
>
> Medin, Ross, & Markman (preface to 'Cognitive Psychology') [26]

1 Introduction

The tremendous progress in artificial intelligence of the past two decades not withstanding [13, 34, 2], our best case for intelligence remains arguably what different minds, in the biological world, achieve. A mind has limited computational resources, in time and space, and we view it as all those information and decision-making processes that are collectively in charge of the difficult task of

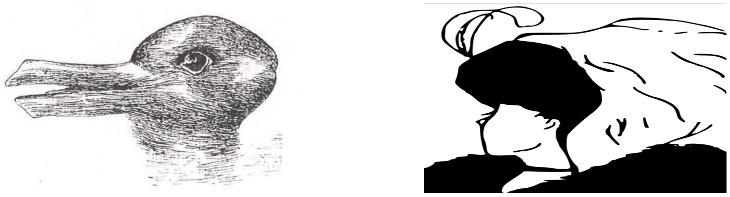


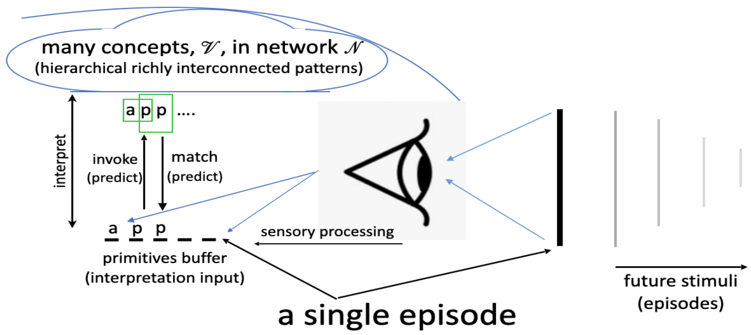
Fig. 1: Subject to interpretation: an image can be interpreted in a variety of ways, and some interpretations could be much more salient to the observer (interpreter) than others. Left: rabbit or duck. Right: old *vs.* young woman (among other possibilities). In interpretation, one's perceptual processes determine which of one's concepts (hierarchical patterns) are present and in which configurations (their spatiotemporal relations). Interpretation answers what (pixels, lines, ..) goes with what, and generating 'perceptual stories'.

conducting the business of life (maintaining, adapting, growth, prolonging, managing). Thus, a mind is *finite* and continually interfaces with a world (including the body) that is, in our view, *infinite*.[1] The external world is infinitely rich, changing, and productive, but devoid of meaning, and minds create or *extract their own meanings* over time by interacting [27, 29, 9, 8, 6].[2] This interaction, observed over some spatio-temporal period, can be interpreted as flexible behavior and called intelligent (by another mindful subject).

The intelligence of humans appears to rest on entities, or representations, called concepts [28, 18, 12] (see Sect. 2.2), and in particular *many concepts*, such as 'mother', 'water', 'table', 'democracy', 'life', 'mind', 'mad', $\cdots$, that work well enough together, for instance to achieve daily common sensical behavior. Two distinct and major problems (among many) that arise are:

[1] Boundedness does not imply finiteness. That the 'physical' universe is finite or infinite is probably not a well-formed provable (or even falsifiable) statement, and an answer to the question of *what* is to be counted would need to be agreed upon (related to a *mind's interpretation*). However, in one perspective, the trajectory, *e.g.* from physics, to chemistry, to organic chemistry, to life and society, points to the world's productivity and open-endedness. We also posit that it is best to think that our minds, due to development, have the potential to interpret a scene, or a snapshot of sensory impressions, in an infinite variety of ways. This paper touches on that interpretation aspect.

[2] The notions of worlds and minds are intertwined. Consider, for instance, this quote, from Lewontin, 1991: "[T]here is no "environment" in some independent and abstract sense. Just as there is no organism without an environment, there is no environment without an organism. Organisms do not experience environments. They create them." (from [3]). See also "lifeworld" of Husserl and 'Umwelt' of von Uexküll, and enactive and embodied cognitive science [17, 40, 41].



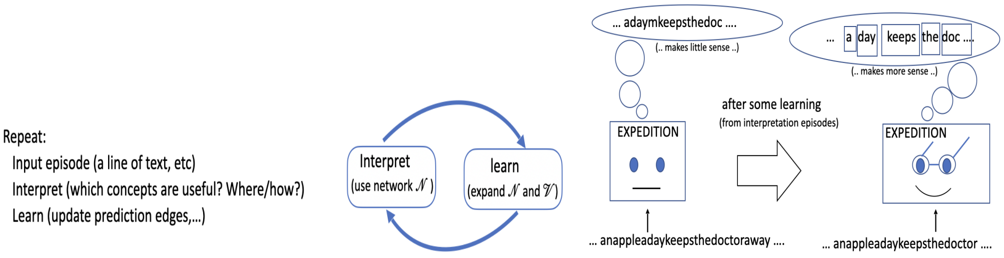
Interpretation: which (few) of my many concepts form a *coherent account* of the input?

Fig. 2: Interpretation in Prediction Games (PGs): In each (interpretation) episode, sensory pipelines convert the raw impressions into a primitive sequence, then interpretation takes place, answering the question: out of my many (millions of) concepts in $\mathcal{V}$, which few, *e.g.* 10s to 100s, are useful in this episode (and in what configuration)? In this picture, bigrams (corresponding to) 'ap' and 'pp' are activated.

1. (a snapshot-in-time question) Given sensory input, for instance when looking at a picture or a scene at time t, how does one, quickly and adequately, figure out (mostly unconsciously) which of one's concepts (patterns), from a large set $\mathcal{V}$, are useful (*i.e.* adequately meet one's needs at time t)? (Fig. 1)
2. (historical/developmental question) Where do these many concepts (in the thousands, millions, ...), richly interdependent, come from, in the first place?

On the question of where so many concepts come from, supervised machine learning offers a candidate answer: that they can be taught (*i.e.* labeling, manually or by some process). However, it appears that this is not feasible, in achieving human-level capabilities, considering the amount of the learning that is required, and the complexity of the interaction between an individual and her/his world. The concepts are to be operational or useful in some sense (they are not merely for classification or labeling). When looking at infant development, it also appears that much learning takes place without explicit teaching or communication: that appropriate machinery, and some appropriate conceptual space(s) (or 'theories' of one(s) world [31, 10]), has developed already in the mind, after the early few months of life, that makes learning from other humans possible, such as learning a mother tongue and further developing appropriate behavior. We seek unsupervised continual learning of many concepts by a system that is embedded in its environment.

The way we have tackled these two questions, and the constellation of problems around them, in our *dynamical systems* approach, which we have called ***Prediction Games*** (PGs) [20, 19], is to address the two questions simulta-


Repeat:
Input episode (a line of text, etc)
Interpret (which concepts are useful? Where/how?)
Learn (update prediction edges,...)
Interpret
(use network $\mathcal{N}$)
learn
(expand $\mathcal{N}$ and $\mathcal{V}$)
... adaymkeepsthedoc ....
(.. makes little sense ..)
EXPEDITION
... anappleadaykeepsthedoctoraway ....
after some learning
(from interpretation episodes)
... a day keeps the doc ....
(.. makes more sense ..)
EXPEDITION
... anappleadaykeepsthedoctor ....

(a) Expedition learns via *practicing* interpretations.

(b) After learning in many interpretation episodes, Expedition interprets with higher-level (more predictive) concepts (n-grams).

Fig. 3: (a) The basic loop of Expedition repeats: input episode→ interpret→learn (from the interpretation) ("practice interpretation"). Learning currently consists of updating prediction edges and, from time to time, adding new concepts (compositions). (b) After many interpreting+learning episodes, Expedition interprets (predicts, etc) with higher-level concepts (with increased information-utility): the same input can evoke different, highest level, concepts in the future.

neously. We imagine the PGs system to be composed of several learning and inference processes working together. Our current system, called **Expedition**, repeatedly *practices using its existing concepts*, as it ***interprets*** its raw input (Fig. 1 and 2), in order to develop its conceptual space further (*i.e.* update and enrich its network of connections $\mathcal{N}$, and from time to time add new concepts, expanding its concept vocabulary $\mathcal{V}$) (Fig. 3). We note that the structured PGs approach, currently focused on developing perceptual concepts (percepts), could allow ultimately for communicating among such systems (language development) and *teachable* systems. Compared to existing neural network approaches, the concepts are explicitly represented and modular (discrete), which could have advantages of transparency (interpretability), and sample efficiency when learning or adapting (when assigning blame or credit) (see Sect. 3). The PGs approach is a model-based framework, but *myriad models* are learned and adapted, in part concurrently, and in part sequentially and cumulatively. It can be viewed as *constructivist* [8, 5], but focused on perception. There remain many challenges, as we explain, but our findings point to the promise of this framework.

This summary paper is intended to provide a short overview of the current ideas and system, with updated motivations and philosophy, and some findings and comparisons. The next section explains the (Expedition) system [21, 22], and goes over basic notions, including what we mean by "concept" or "interpretation". We then summarize our findings to date. In Sect. 3 we compare and contrast the PGs approach, conceptually, to large language models based on neural networks, and in Sect. 4 we describe possible applications of PGs.

2 Overview of the System and Basic Notions

Expedition works on text and the concepts it develops correspond to n-grams of characters currently (Fig. 2 and 3).[3] At any point, the system maintains a vocabulary of concepts $\mathcal{V}$ and a hierarchical network of (typed, prediction) edges $\mathcal{N}$ among them. Both sets are dynamic: $\mathcal{V}$ grows, and edges are added and dropped from $\mathcal{N}$ during the learning. Initially, $\mathcal{V}$ is limited to the set of ***primitive*** concepts,[4] corresponding to an alphabet (a discrete finite set), $\mathcal{A}$, and initially $\mathcal{N}$ can be empty (or $\mathcal{V}^{(0)} \approx \mathcal{A}$ and $\mathcal{N}^{(0)} = \{\}$ at time $t = 0$). In text, $\mathcal{A}$ can be the set of ASCII characters, up to 100 unique characters in our experiments. However, we have also experimented with lower level primitives, which we explain later.[5] Expedition learns by repeatedly inputting a text string, such as a line, randomly picked from a text source (a text corpus, or a large file), interpreting it, and learning from this process (an interpretation episode): $\mathcal{N}$ and $\mathcal{V}$ grow over time. For example, in our experiments, over say 100s of thousands of episodes, $\mathcal{V}$ grows from ≈100 to 10s of thousands of concepts.

2.1 Interpretation

There is a one to one correspondence between a character (from alphabet $\mathcal{A}$) and a primitive concept. We often say "character", an element of $\mathcal{A}$, and "primitive" interchangeably, thus initially $\mathcal{V} = \mathcal{A}$ (abusing notation). Similarly, we will use 'a', to refer both to the unigram character as well as the primitive concept corresponding to it.

The content of the input buffer, the character sequence, in each episode, is the ground-truth, "reality", for the system, and what the system uses to verify the higher-level concepts it activates. The raw contents are also the starting point of the analysis in an episode, the start of the *interpretation process*. Interpretation is the process of mapping the characters, the lowest level observations, into highest-level concepts in current $\mathcal{V}$ (n-grams). In general terms, it is the process

[3] The support for approximate matching, Sect. 2.4, extends the representation somewhat, to a conjunction of stochastic disjunctions. We conjecture that there exist learning algorithms to further increase the representation power.

[4] These are the 'givens' or the innate or hardwired concepts (given by evolution, or the engineer designers). In Uexküll's work in theoretical biology, the primitives may stand for (the digitized versions of) the inner/primal features (innenwelt or unmediated reality), while higher-level concepts would correspond to the more distal/transformed/mediated features of the umwelt, learned by an organism. Much work, *e.g.* in computer vision (such as recognition by parts) [15, 33, 44], shares similarities to PGs. Our approach is also reminiscent of Hume's atomism (as well as axiomatic mathematics), that perception could be reduced to "atomic impressions" (indivisible building blocks) via various constructions.

[5] On the other hand, in the first implementations of PGs, initial $\mathcal{V}$ was the set of all words [19]. It is not required that the entire $\mathcal{A}$ be specified from the outset, just that its members be automatically recognized (by earlier sensory pipelines) and, whenever a new member is observed (first time), a new primitive be allocated.

of figuring out *which (few) of ones (high-level) concepts are most useful for the current situation*. More specifically, in our approach, interpretation is a search for *the concepts that form a most coherent account of the input*.

In every episode, the system tries to find a sequence of concepts in $\mathcal{V}$ that best 'hang-together' (cohere) as well as 'explain' or match (or predict) the contents of the input buffer (see CORE, Fig. 4(a)). The final product of this process, which we also refer to as an interpretation, is a data structure comprised of a sequence of concepts, and for each concept c an specification of the consecutive characters in the input buffer that concept c matches (accounts for).

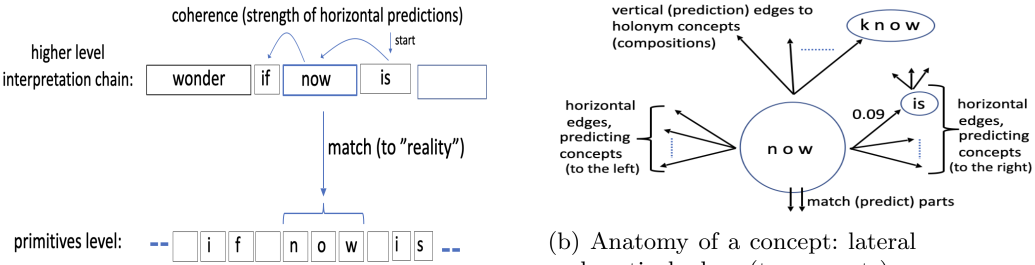
There can be multiple selected interpretations to cover the entire buffer, and there can be overlaps among concepts, in the regions that they account for (but no double rewarding in computing the CORE score). These final selected interpretation data structures are used for learning (weight updates). As higher level concepts (compositions: bigrams, trigrams, ...) are built over time, interpretation becomes a more involved process, consisting of both 'bottom-to-top' and horizontal edges, invoking compositions and other concepts, as well as 'top-to-bottom', or matching, sub-processes: for certain search-initiating concepts (primitives at the lowest level), the system decides whether to join it to its left, or right, or leave it as is, and repeats this, until no well-matching concepts remain. Both invocation and matching are prediction processes. There is evidence that in the human visual system both bottom-up and top-down (prediction) processes are at work during perception [39, 38, 32].

Interpretation also plays a major role in the field of semiotics, the study of sign systems and meaning making [7, 16, 14], and our work can be viewed as a computational (and learning/developmental) account of semiosis.[6]

2.2 What is a Perceptual Concept?

A concept in our current system corresponds to an n-gram, a sequence of one or more consecutive characters (Fig. 4(b)). It can be viewed as a pattern with a *control structure*: once activated, it takes over and directs the matching (or predicting) during the interpretation process. For instance, imagine that the "now" concept exists in the system. It is composed of the primitives 'n', 'o'

[6] Semiotics (*e.g.* biosemiotics and cognitive semiotics) studies how signs and symbols convey meaning (and is not just limited to intentional communication) [7]. Interpretation, in particular in the work of Charles Peirce, is a central concept in semiosis, and there are similarities to our use: the system or agent interprets (maps) a stretch of raw sensory information, internally, into an *interpretant*, which could correspond to our highest level selected concept. For us, the final product of an interpretation process can be a sequence of concepts, and even several interpretations (*e.g.* to cover the entire input buffer), and we use and develop prediction and probability semantics for what could be meant as relations between the semiotic concepts of 'signs', 'icons', and 'indices'. Previously, we used the term 'segmentation' [21, 20], but interpretation is a better fit. Our work, with its 'inward' emphasis on bottom-up concept building for internal prediction and use, is also closer to idealism (*vs.* realism) in epistemology (*e.g.* see [36, 37]).



(a) CORE (COherence + REality).

(b) Anatomy of a concept: lateral and vertical edges (to concepts).

Fig. 4: (a) CORE is used to guide the search and score candidate interpretations during interpreting. Two ways of understanding CORE: 1) a measure of information gain, 2) a way of combining fit to context and match strength (match to the buffer contents). (b) A concept, in $\mathcal{V}$, is a node in the network $\mathcal{N}$ (of typed weighted directed edges). For example, 'now', is part of 'know' and other compositions, *i.e.* it has vertical edges. It also has horizontal edges (to nodes/concepts), each weighted (moving conditional probabilities), for predicting its immediate left and right in interpretations it occurs in.

and 'w', in a certain order or configuration.[7] If 'n', 'o', and 'w' appear in the input together, and the concept "now" fits the wider context well, *i.e.* nearby concepts predict it well or vice versa, it will participate as part of the final selected interpretation (*e.g.* consider ".. the time is now.." *vs.* 'now' appearing in ".. the knowledge .."). Approximate matching is supported too (see Sect. 2.4). Thus, while the canonical form is 'now', the presence of this form in the input is neither necessary nor sufficient for the concept "now" to be picked in a final chosen interpretation of an episode.

A concept is also a node in network $\mathcal{N}$: it has weighted prediction ('horizontal') out-edges and holonym or vertical edges (Fig. 4(b)). Concepts are *both the predictands* (targets of prediction, by other concepts) and the *predictors* (features) in the system. This symmetry is an attractive draw of PGs.[8]

Concepts *vs.* Interpretations. Interpretations, composed of concepts, are ephemeral in general, lasting for an episode only. If episodic memory were supported, perhaps some interpretations could be recorded to serve as useful mem-

[7] The matching order need not be left to right. For instance, in a bottom-to-top search process, the middle 'o' may first activate and start the match/prediction attempt. This allows for flexibility when the entire concept does not appear in the input buffer, due to timing (of what went into the buffer), noise and corruption, and so on.

[8] This symmetry (and constraint) is not typical in machine learning, for instance, in supervised learning (nearest neighbors, decision trees, neural networks, ...), the features (or predictors) are often different from the classes (the predictands).

ories. Interpretations can be viewed as (fleeting) micro-thoughts, or "perceptual stories" (Fig. 1). Concepts can be viewed as pieces of interpretations that became persistent at some point (Sect. 2.5), lasting across many episodes and perhaps for the entire remainder of the life of the system. Concepts collect and update (prediction) statistics and are utilized in a slow (statistical/associative) learning process.

2.3 Edge Weight Semantics: Moving/Developing Probabilities

Concepts predict one another, and these prediction relations are supported by weighted directed edges. The weights are conditional probabilities that change over time. Expedition supports predicting what comes immediately to the left and to the right of a concept, in a final selected interpretation.[9]

At the primitives level, when higher level concepts haven't been learned yet (when $\mathcal{V} = \mathcal{A}$), the edge probabilities have clear simple semantics: we have two sets of prediction edges, the left and the right, and for instance for the right edge, 'a' → 'b', the weight is the conditional probability, $P('b'|'a')$, that 'b' follows 'a' (immediately to the right of 'a') in the input buffer.

Non-Stationary (and Non-IID). With the generation and use of new higher-level concepts in interpretations (the growth of $\mathcal{V}$), the conditional probabilities, in addition to being dependent on the (external) input stream, become a function of the system's decisions in the previous interpretation episodes, such as which interpretations won and were selected (the system's historical behavior). The same holds true of new concepts generated.[10] The input stream, to any concept (serving as a predictor), is non-IID. In particular, the use of new concepts causes non-stationarities, even if the external world is stationary. Certain weights (probabilities) need to be lowered (or edges possibly dropped altogether), while new edges are introduced or probabilities raised. This is a kind of *internal*, and in particular *developmental*, non-stationarity: the external input stream (text corpus, etc) need not be changing, but internally the system goes through changes. We investigate learning and tracking changing probabilities via space bounded predictors, which we call Sparse Moving Averages (SMAs) [23].[11] For this type of open-ended non-stationarity, challenges of plasticity *vs.* stability arise, and to

[9] Learning longer predictions are possible, and in earlier implementations of PGs we used a larger window, but, currently, CORE works adequately with simple 'one-hop' prediction.

[10] This is a second sense in which Expedition is a "system" in a *dynamical-systems-theory* sense: the initial conditions and its interpretation history affects its future behavior (interpretations) through self-feedback, *i.e.* edge-weight updates. This self-feedback phenomenon becomes more pronounced when PGs learning is embedded in a larger agent that decides what to sense (see Sect. 4). This dependence may make theoretical analyses more challenging as well (perhaps to a lesser extent, other learning techniques, in particular the online ones, share the same property).

[11] Sparse, as $\mathcal{V}$ is large, but for an update of an edge only one concept is observed at any time point. The number of edges (prediction relations) is kept sparse too, *e.g.* in

best handle such, different learning-rates, one for each edge (prediction relation), are kept. These SMA techniques are used for updating the active concepts' prediction edges (whenever a final interpretation is selected). This is a concurrent online aspect of the learning: multiple edges, but a relatively small subset of the entirety of $\mathcal{N}$, are updated in each episode.

Thus, in PGs, the edges in $\mathcal{N}$ may encode not just weights (conditional probabilities), but also learning rates and other information useful for, in effect, change detection [23], as well as other match-related information (*e.g.* concept offsets). The search process during interpretation uses the edge probabilities for a more effective search, as well as for scoring and selecting final interpretations (see CORE next). We are working on developing probabilistic semantics for the vertical edges as well, to make the search more robust and principled.

2.4 CORE: Why Should I Compose? (Gain in Information)

Intuitively, 'now' is more informative or meaningful (in the English language), than any of its individual parts, such as 'n' or 'o'. CORE quantifies this in terms of a comparison to a simple baseline prediction system [22]. If a composition such as 'now' is predicted by the system (context, *i.e.* concepts nearby, or a prior) with sufficiently high probability p, in particular if p is larger than the product of the character-level probabilities, p_i, of 'n', 'o' and 'w' (the primitive priors), then it is better to predict 'now' (compared to predicting the individual characters separately). CORE is $\log(\frac{p}{p_1 p_2 p_3})$, and it is a *gain in information*[12] over a baseline system that (1) does not go beyond individual characters, and (2) makes the independence assumption (does not even learn the conditional probabilities among the unigrams). Expedition, by expanding $\mathcal{N}$ and $\mathcal{V}$, gets more and more distant from the baseline in terms of CORE (when run on a non-random stream) [21, 22].

CORE is flexible and supports approximate matching with a principled appropriate penalization, *e.g.* substituting 'n' in 'now' with any string in $\mathcal{A}^*$ (Kleene closure) (such as the possibility of skipping 'n'). Fig. 4 shows that CORE can also be viewed as a way of combining strength of fit to context with strength of a match (each with its own costs and rewards, all based on logarithm of probabilities). We use CORE to guide the search and score interpretations (a chain of concepts). There are a number of ways that CORE could be used in the search, and we continue to investigate how best to utilize it. For further details about CORE, please see our prior work [22].

2.5 Challenges of Generating and Incorporating New Concepts

There are a few common themes, for generating and incorporating new concepts, and our overall approach so far could be viewed as an implicit generate-and-test

the 100s. This implies good probabilities can be learned (tracked) down to a certain minimum (*e.g.* 0.01 or 0.001) [23].

[12] The term "information gain" has another somewhat different meaning in decision tree induction [30], but its use is also very applicable here.

scheme. In general, the system cannot predict whether or how much CORE is improved by a candidate new concept: this depends on how different the distribution around the composed concept is, compared to its part concepts (is there synergy?). The system generates a number of concepts over time, and let them gather statistics and in effect cooperate and compete, with existing and other new concepts. If there are higher level regularities in the input stream, CORE should eventually improve.

The (horizontal) prediction edges are good candidates to generate holonyms (compositions) from. A possibly simpler alternative is generating, with some probability, a composition from two concepts next to each other in an interpretation. However, a concept needs to be *used (observed)*, *i.e.* selected in sufficiently many final chosen interpretations, before it is allowed to participate in compositions (or allowed to produce 'children' nodes or holonyms of its own). Otherwise, the system could generate too many (insignificant or low utility) concepts. Typically we have required 100s of observations as a threshold. Another problem surfaces: a new concept may not have any horizontal (out-going) edges when it is created, and other concepts are not predicting it either (no incoming edges). But such edges are required to have a good CORE score to be selected (to be "observed"). This is a *chicken-and-egg problem*,[13] and we are investigating interpretation processes that can accommodate the robust incorporation of new concepts. It is possible that some form of balancing between exploration and exploitation is required.

Composing to generate higher n-grams can be viewed as a kind of *coarse-to-fine* generation (from general such as a primitive 'a', to more specific such as 'apple').[14] We expect going in the other direction, *i.e.* supporting a kind of grouping or clustering (disjunction or compression, or fine-to-coarse) in the structure of concepts is also fundamental and crucial to achieving the full power of PGs. This remains a major open problem.

2.6 Some Findings

In our earlier implementation, in order to deal with the non-stationarity caused by generation of new concepts, in order to incorporate, the Expedition system had to learn from scratch every so often [21]. With advances in non-stationary probability prediction (SMAs) [23], and how we use CORE, the system is significantly simpler and these transitions (incorporating new concepts) are smoother.

As the system learns, CORE improves over time [21, 22],[15] and the concepts discovered appear to correspond well to words and phrases. Furthermore, one

[13] Recommender systems have a similar 'cold-item' problem for incorporating, that is appropriately recommending, new items.

[14] Another, more common coarse-to-fine operation (such as in inducing hierarchical clusterings or in a few techniques for learning finite-state automata) is to split an existing node into two or more nodes (currently missing in Expedition).

[15] However, related information theoretic measures such as entropy or perplexity degrade as the vocabulary $\mathcal{V}$ expands.

can examine the final interpretation in an episode and examine how the system 'sees' that episode (it is an 'interpretable' approach, see Sect. sec:apps).

Measuring Quality of the Endings. Expedition makes local decisions, and its interpretation search needs to be fast, so it is possible that, for instance with inferior algorithms, it could converge and get stuck in poor local optima. It would be good to have quantitative external measures of progress (CORE being an internal measure). We note also that the learning of new concepts is not uniform, *e.g.* some trigrams are built before discovering all good bigrams (and common phrases discovered before many single words). We have been using measures based on goodness of the endings: how well the endings of concepts, in an interpretation, respect the "natural" but hidden (to the system) boundaries. As larger concepts are discovered and used, do we get better alignments? In natural language, if we take blank space and other punctuation as good places for the endings, and inside of words as bad ones, indeed the system, as it learns, reduces the bad-endings ratio (*e.g.* the ratio of bad right-endings to all right-ends in an interpretation) from say 0.7s, *i.e.* 3 out of 4 splits are bad (when it begins with unigram characters), to 0.2s.

We have observed that more search time allowed within each interpretation episode can help improve CORE in the episode, as expected, as well as bad-ratio performance [21], but more research is required to understand the interaction of various components, and, for instance, whether (and how) concept structures should be modified (currently, only prediction edges are modified).

Encoding Experiments: Assessing Dependence on the Choice of Primitives. As the vocabulary $\mathcal{V}$ grows, the new concepts appear less frequently, and gathering statistics and the learning slows in a sense. We want to evaluate repeated compositions (and whether/how errors accumulate) quicker if possible, and also assess dependence of performance on the quality (informativeness) of the starting alphabet. Thus, we are also conducting experiments where Expedition begins with an alphabet $\mathcal{A}'$ set to a lower level than ASCII characters,[16] for instance, at an extreme, binary $\mathcal{A}' = \{$'0', '1'$\}$ (the system would have two primitives only). We encode each original character in $\mathcal{A}$, via a fixed or variable-width encoding, using codes (strings) based on the elements of $\mathcal{A}'$ (Fig. 5). How well are the hidden character boundaries discovered, when Expedition starts with such $\mathcal{A}'$? An early such experiment was reported in Expedition [21], with some positive results, which we have improved since. For instance, when using fixed-width bigram encoding, *i.e.* $|\mathcal{A}'| = 10$, yielding 10x10=100 codes (adequate for up to 100 unique original characters in $\mathcal{A}$), and initially every other ending would be bad at the primitives level (Fig. 5, baseline of 50%). Our current techniques lower the bad-ratio to below 0.04 rate, within a few 1000s of episodes.

By stressing the Expedition systems in these ways, we explore the various tradeoffs involved and get insights into the limits of the learning algorithms, and also find opportunities to improve them.

[16] Thanks to Brian Burns for pointing out the possibility of intermediate alphabets.

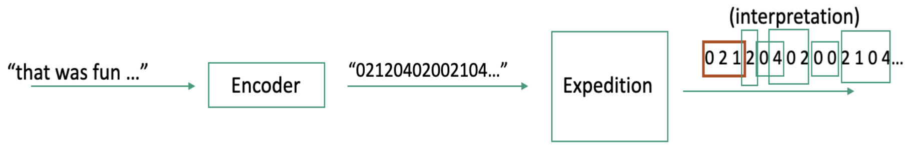


Fig. 5: In an encoding experiment, each original character gets a code from a lower level alphabet $\mathcal{A}'$, *e.g.* here $\mathcal{A}' = \{$'0','1', ...,'9'$\}$. In the example above, each character is encoded into a bigram (fixed-width 2) from $\mathcal{A}'$ (*e.g.* 't'→ '02', 'h'→ '12',..). With Expedition starting with $\mathcal{A}'$ ('sees' its input stream at the level of $\mathcal{A}'$ initially), the general question is how well it can recover higher-level (hidden) regularities. After many episodes, interpretation is performed at higher levels (bigrams, trigrams, ..., over $\mathcal{A}'$), and one can evaluate the left or right-endings of the activated concepts in each interpretation. In such fixed-width bigram settings, half of the endings (every other one) are bad (50% bad-ratio for a baseline performance, and the initial performance of Expedition), and Expedition, with our current techniques, reduces its bad-ratios to nearly 0 (below 0.04) after a few thousand episodes. In the Expedition output above, the first ending (in red) is bad, and the remainder (in green) are good.

2.7 A Summary of the Main Assumptions

We summarize a few main interrelated assumptions (or informed intuitions) behind the PGs approach:

- (discreteness) A discrete approach to concepts, to their representation, is feasible (could this hold true only for perceptual patterns?).
- (sparsity) Not many (out) edges needed per concept (*e.g.* 100s or 1000s).
- (approximate) No need for learning probabilities below a certain minimum.
- (fresh data) The rich unlimited input stream is an important enabler (for feasibility).
- (local/online) Making local decisions (*e.g.* in an interpretation search) is adequate for robust learning (*vs.* expensive global optimization).
- (algorithms) There exist incremental unsupervised techniques for learning sufficiently powerful representations (and learning many such patterns).
- (a system, not a single algorithm) Multiple processes, of learning and inference, working together, is probably required (and some may need to be tightly integrated).
- (one objective for PGs) Prediction suffices as an overall driver of the learning.
- (open-endedness) Having no a priori size limit on $\mathcal{N}$ is a considerable draw.

On the main feasibility assumption, the persistent question is whether we can avoid combinatorial explosion while preserving the learning of powerful representations. We have seen promising results combining SMAs for prediction edges, interpretation search and inference techniques using CORE as the guide, and basic concept generation and incorporation ideas [22, 23]. Much work remains to be done to provide evidence for truth status of the above. In particular, progress on large deep feedforward neural networks, *e.g.* large language models based on transformers, cast doubt on some, which we discuss next.

3 Conceptual Comparisons to Large Language Models

Large language models (LLMs), based on feed-forward neural nets (NNs) in particular using the transformer architecture [1, 42], are also based on learning and operating via predicting, and have attained substantial success, surpassing all expectations [42, 13, 2]. We have motivated PGs along the same lines [19, 20], but as discussed above, there remain many research challenges. What could PGs offer? We review some conceptual differences between LLMs and PGs, and point to a few potential benefits, and disadvantages, of PGs next.

Supervised vs. Unsupervised. NNs are supervised learners, requiring pairs <x, y> (a feature vector, or token sequence, x, and a label or class y), while PGs are unsupervised (self-supervised) structure learners (no use for the y's): a PGs system makes its own y's (concepts, or functions). The learning is open-ended: no a prior bound on network size[17] (unlike a typical NN, whose finite structure needs to be specified). To function, in each episode, the PGs system first structures the input stretch x, imposing its own current concepts and their learned relations, onto x (it interprets). This allows it to make predictions, with (structured) concepts as the predictands and predictors, of what may come next for example. This *unsupervised learning* can be useful for environments/tasks with little domain knowledge (see next section).

Cost of Inference. During operation, even if we stop the learning, PGs require two-way inferences, involving search, activating, matching, and scoring (Sect. 2.1). While we work on advancing and simplifying the algorithms, interpretation remains a costly elaborate process that, for instance, requires data structures that keep account of the matchings attempts,[18] and each concept (node in $\mathcal{N}$) can be viewed as a sophisticated 'book keeper' for predicting changing probabilities in an open-ended setting [23, 22].[19] NNs, are significantly lighter, using one-directional inferences (and a node in the network is much simpler). Thus operating PGs could be costlier.[20] What could the learning of structures and predicting with them buy us?

Sample Efficiency (or Agent-Friendly Learning). NNs are powerful function approximators, and currently PGs learn n-grams only. We conjecture

[17] This is in theory, or the abstract model of the learning: the network size $\mathcal{N}$ in PGs is a function of experience.

[18] Techniques such as cachings of various kind (from recent episodes) could help substantially, but the cost of doing good search and inference cannot totally be erased.

[19] Even the edges may need to encode not just weights, but other information, such as learning rates, match offsets, and change detectors (see Sect. 2.3) [22, 23].

[20] However, many NN layers may be required to achieve similar functionality. Empirical comparisons are needed ultimately.

that the representation power can be extended, but we expect some bias, *a limitation*, in learning to remain. However, this bias may be a benefit in terms of *sample efficiency.* In particular, even if we assume each (perceptual) learning episode, say in a human, takes place in as low as the order of a second, humans have only about 10^6 episodes of learning in a year. The issue of limited experience is exacerbated when one considers action execution (actions can take more time than perceptions, may change environment in an irreversible way, etc). NNs are in general sample hungry, and may require more samples for the extensive complexity of what humans learns. Furthermore, the world changes, and for instance skill learning may need to be continually tuned and adapted (one sense of *dynamic coupling* with the environment [25, 4, 41, 6]). The representation in NNs are distributed and often very deep (to achieve their power). It is hard to update one aspect without changing others, the extreme version of this being issues of catastrophic forgetting in sequential learning [43]. For such reasons, NN training typically requires IID samples (randomly permuted corpora, with many training passes over such). PGs use more of a localist (digital) representation [11]: while there is a distributed aspect, *i.e.* higher level concepts can share part concepts, each concept acquired is explicitly represented (a node in $\mathcal{N}$) and explicitly participates in interpretations. Updating in PGs is more localized (the edges of those in a final selected interpretation are updated), and higher level concepts, being less frequent than the lower level, take more time to learn and thus tend to be substantially shallower than typical deep NNs.[21] Thus, PGs may incur a higher computational cost for interpreting and learning in each episode, but may incur a lower sample size cost (trading learning complexity for computational complexity).

The SMA techniques, utilized in PGs, are designed for efficient learning under internal and external non-stationarity [23, 22] (Sect. 2.3), and thus PGs may be more robust toward sequential non-IID learning in general. The next section expands on this.

4 Potential Applications

We hope research in PGs can shed further light on processes involved in perception and development (as a theoretical and computational framework). Other applications of PGs can be broken into two categories: stand alone, *e.g.* as a learning tool for analysis, and in a larger system, *e.g.* in an (autonomous) agent.

An Interpretable Learning Tool. As a stand alone, a PGs system provides an interpretable tool for analysis, *e.g.* of a foreign language or for a recorded unknown activity (animal communications and activity, computer network behavior, and so on). We assume the primitives to be a set that conveys some

[21] Considering these observations, the concept of "development" does not readily apply to deep NN training. However, in the case of PGs, one can look at the transitions to using new concepts (very identifiable) and the changes in the prediction probabilities, as development (*i.e.* a series of changes, forming trajectories in the learning tasks).

initial understanding by the analysts,[22] and that sufficiently many episodes are available (a long enough data stream). Some locality of dependencies (regularities) needs to hold too, but we expect many natural and artificial systems to follow such hierarchical but local patterns, or the near decomposability of complex systems [35]. A concept acquired is interpretable, it has a canonical form, a sequence of primitives currently. Its summary context is also reflected in the system, *i.e.* the lateral prediction edges with their probabilistic weights: in general, a node's neighborhood in $\mathcal{N}$. Thus, one can examine the network $\mathcal{N}$ learned. An interpretation itself is interpretable: one can examine how the system 'sees' its input (in a given episode), by looking at high scoring interpretations.

Agent-Friendly Learning. In a larger agent (in which PGs could be a subsystem), the higher-level concepts learned, *e.g.* discovered and extracted in the visual modality, can then find a variety of applications, serving as more powerful predictors, into more distant future/space, than the lower level primitives, in the same modality or data stream, or for a different modality, *e.g.* better predictors of food, and reward/cost in general (associations within and across sensory channels). The kind of perceptual experience an agent receives as it acts is biased toward certain aspects of the environment, such as finding food, shelter, and so on. We expect PGs can be more systematic and faster online learners than traditional NNs and to be more effective at generalizing, in a world that, in many senses, is highly structured but also changing in significant ways. Finally, PGs can provide the (random) discrete-valued variables (the activated concepts in an episode), along with appropriate probabilities, that are necessary for reasoning and planning, communication and language (and teaching), episodic memory, and other symbol-based tasks under uncertainty. We seek to explore these possibilities within autonomous agents [24].

5 Conclusions and Future Directions

Perception is not a simple input interface: it may be viewed as the point at which a finite (but dynamic and growing) mind interfaces with an infinite world. Through perception, at every time point, the mind determines which few of its many concepts are the variables of interest/concern (repeatedly answering the question: what should be my concern?). The world is infinite and does not directly provide this: the mind (agent) has to identify and extract such [29, 9, 27]. With development, the set of possibilities (concepts) can change and expand. PG learners are designed for such changing and open-ended internal and external interactions. A PGs system is not given a priori what high-level complex patterns (concepts) will be useful in the world it will face, but is provided with a set of building blocks and appropriate machinery and biases (learning and inference algorithms), and has to discover and incorporate such patterns in a timely developmental, *i.e.* sequential and cumulative, manner. In particular, it *practices interpretation* and through such learning it continually adapts and expands its

[22] Concepts learned on an initially low level, such as the binary primitives of Sect. 2.6, may take some time to understand.

network of concepts. We seek to incorporate PGs within autonomous agents, and explore how the PGs approach could support other types of learning, for instance in foraging tasks and in cumulative skill learning [24].

How far can one extend the learnable expressive power of PGs? A major open direction is extending what can be learned to well beyond n-grams, *i.e.* beyond *conjunctions*: in particular supporting learning and incorporating in effect *disjunctions* together with conjunctions. We may need to extend the current scoring function, CORE, in order to do so. Formally, if the episodes are generated by a (probabilistic) finite state machine (FSM), and the input buffer is large enough to contain every string output by the FSM, we conjecture that CORE is maximized if the FSM is learned. This can be established when the FSM is restricted to a subclass (for instance, when limited to n-grams). Furthermore, it would also be good to develop guarantees of convergence to the underlying FSM with appropriate assumptions and (PGs) algorithms.

Finally, related to the above questions of learning power, given sufficient agreement on the basic philosophy and motivations behind the PGs approach, another broad question is the number of distinct *hardwired (basic) processes* that one needs to approach human capabilities (engineering complexity).

Acknowledgments

Many thanks to our weekly SAIRG reading and research group (including: Tom Dean, Brian Burns, Reza Eghbali, Georgi Georgiev, and Christian Martinez), for insights and encouragements, to the reviewers for valuable feedback, and to the BICA organizers.

References

1. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, 2014.
2. T. B. Brown and et. al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
3. F. Capra and P. L. Luisi. *The Systems View of Life: A Unifying Vision.* Cambridge University Press, 2014.
4. E. A. DiPaolo, T. Buhrmann, and X. E. Barandiaran. *Sensorimotor Life: An enactive proposal.* Oxford University Press, 2017.
5. G. L. Drescher. *Made-up Minds - a constructivist approach to artificial intelligence.* MIT Press, 1991.
6. H. L. Dreyfus. Why Heideggerian AI failed and how fixing it would require making it more heideggerian. *Philosophical Psychology*, 2007.
7. C. Emmeche and K. Kull, editors. *Towards a Semiotic Biology: Life is the Action of Signs.* Imperial College Press, 2011.
8. C. T. Fosnot, editor. *Constructivism: Theory, Perspectives, and Practice.* Teachers College Press; 2nd edition, 2005.
9. S. Franklin. *Artificial Minds.* MIT Press, 1995.
10. A. Gopnik and A. N. Meltzoff. *Words, Thoughts, and Theories.* MIT Press, 1997.

11. C. G. Gross. Genealogy of the grandmother cell. *The Neuroscientist*, 2002.
12. D. R. Hofstadter and E. Sander. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking.* Basic Books, 2013.
13. E. Horvitz and T. M. Mitchell. Scientific progress in artificial intelligence: History, status, and futures. In A. Mazza, W. Kearney, and KH Jamieson, editors, *Realizing the Promise and Minimizing the Perils of AI.* U. of Penn. Press, 2024.
14. P. Konderak. *Mind, Cognition, Semiosis: Ways to Cognitive Semiotics.* Lublin, Polska: Maria Curie-Sklodowska University Press, 2018.
15. N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. H. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
16. K. Kull. Choosing and learning: Semiosis means choice. *Sign Systems Studies*, 2018.
17. K. YH Lagerspetz. Jakob von uexküll and the origins of cybernetics. *Semiotica*, 2001.
18. G. Lakoff and M. Johnson. *Philosophy in the flesh: the embodied mind and its challenge to western thought.* Basic Books, 1999.
19. O. Madani. Prediction games in infinitely rich worlds. In *AAAI Fall Symposium*, 2007. Yahoo! Research Technical Report, available at: www.omadani.net/publications.html.
20. O. Madani. Systems learning for complex pattern problems. In *Biologically Inspired Cognitive Architectures (BICA) at AAAI Fall Symposium Series*, 2008.
21. O. Madani. Expedition: A system for the unsupervised learning of a hierarchy of concepts. *ArXiv*, 2021.
22. O Madani. An information theoretic score for learning hierarchical concepts. *Frontiers in Computational Neuroscience*, 17, 2023.
23. O. Madani. Tracking changing probabilities via dynamic learners. *arXiv*, 2024.
24. O. Madani, B. Burns, R. Eghbali, and T. Dean. When remembering and planning are worth it: On navigation strategies under change. *To appaer in BICA*, 2025.
25. H. Maturana. Autopoiesis, structural coupling and cognition: A history of these and other notions in the biology of cognition. *Cybern. Hum. Knowing*, 2002.
26. D. Medin, B. Ross, and A. Markman. *Cognitive Psychology.* Wiley, 2000. Quote from the preface.
27. M. Merleau-Ponty. *The Phenomenology of Perception.* Gallimard, 1945.
28. G. L. Murphy. *The Big Book of Concepts.* MIT Press, 2002.
29. S. Oyama. *The Ontogeny of Information.* Duke University Press, 2000.
30. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1986.
31. D. H. Rakison and L. M. Oakes, editors. *Early Category and Concept Development: Making Sense of the Blooming, Buzzing Confusion.* Oxford University Press, 2003.
32. P. Rossel, C. Peyrin, A. Roux-Sibilon, and L. Kauffmann. It makes sense, so I see it better! contextual information about the visual environment increases its perceived sharpness. *Journal of experimental psychology. Human perception and performance*, 2022.
33. Z. Si and S. Zhu. Learning AND-OR templates for object recognition and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
34. D. Silver and A. Huang et. al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.
35. H. A. Simon. *The Sciences of the Artificial.* The MIT Press, third edition, 1996.
36. B. C. Smith. *On the origin of objects.* Bradford Books, 1996.

37. B. C. Smith. *The promise of artificial intelligence: reckoning and judgment.* MIT Press, 2019.
38. C. Teufel, S. C. Dakin, and P. C. Fletcher. Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, 2018.
39. C. Teufel and P. C. Fletcher. Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, 2020.
40. E. Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Harvard University Press, 2007.
41. F. J. Varela, E. Thompson, and E. Rosch. *The embodied mind, revised edition: Cognitive science and human experience.* MIT press, 2017. (revised edition).
42. A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeuroIPS*, 2017.
43. L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
44. L. Zhu, Y. Chen, and A. L. Yuille. Recursive compositional models for vision: Description and review of recent work. *J. of Mathematical Imaging and Vision*, 2011.