Decoding Eye Movements in Cross-Situational Word Learning via Tensor Component Analysis

Andrei Amatuni, Chen Yu
{aamatuni, chenyu}@indiana.edu
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St. Bloomington, IN 47405 USA

Abstract

Statistical learning is an active process wherein information is actively selected from the learning environment. As current information is integrated with existing knowledge, it shapes attention in subsequent learning, placing biases on which new information will be sampled. One statistical learning task that has been studied recently is cross-situational word learning (CSL). In CSL, statistical learners are able to learn the correct mappings between novel visual objects and spoken labels after watching sequences where the two are paired together in referentially ambiguous contexts. In the present paper, we use a computational method called Tensor Component Analysis (TCA) to analyze real-time gaze data collected from a set of CSL studies. We applied TCA to learners' gaze data in order to derive latent variables related to real-time statistical learning and to examine how selective attention is organized in time. Our method allows us to address two specific questions: a) the similarity in attention behavior across strong vs. weak learners as well as across learned vs. not-learned items and b) how the structure of attention relates to word learning. We measured learners' knowledge of label-object pairs at the end of a training session, and show that their real-time gaze data can be used to predict item-level learning outcomes as well as decode pretrained item knowledge.

Keywords: cross-situational word learning, factor decomposition, selective attention, statistical word learning,

Introduction

The everyday world is a complicated setting for learning words. Whenever a word is heard by a learner, there are usually many potential referents present at that moment. This presents a real problem for word learning, as inferring which labels go with which objects is inherently ambiguous. Quine (1960) famously framed this as a problem of *referential uncertainty*. A recently proposed solution to this problem is termed cross-situational learning. In cross-situational word learning (CSL), learners don't have to infer the correct referent for a word within one learning situation. Instead, they integrate statistical evidence across multiple learning situations to reduce uncertainty about the correct label-object mappings. Experimental studies have shown that both infants and adults can successfully map novel words to their visual referents cross-situationally, under varying degrees of referential uncertainty (Smith & Yu, 2008; Yu, Zhong, & Fricker, 2012).

In most statistical learning paradigms, such as those used in CSL experiments, researchers encode certain statistical regularities into training stimuli and measure whether learners can successfully use those regularities to infer new knowledge. Prior CSL studies have demonstrated a number of different learning effects by manipulating stimuli statistics showing, for example, that Zipfian frequency distributions for label-object pairings may actively aid learning (Hendrickson & Perfors, 2019), that 16 month olds prefer massed pairings versus conditions that space presentations evenly in time (Vlach & Johnson, 2013), and that children learn better in conditions where label-object pairs are embedded within diverse contextual frames (Suanda, Mugwanya, & Namy, 2014). These sorts of statistical features of the input (e.g. the shape of the frequency distribution or the temporal/contextual statistics) drive learning in different directions. However, even in cases where subjects demonstrate successful learning, it's unlikely that they keep track of all the regularities encoded in the training stimuli. Instead, human cognitive systems are selective. Learners take an active role in selecting their learning curricula from the set of available statistics, with different sampling schemes leading to individual differences in learning performance. Prior work has shown that fine-grained differences in attention behavior predict learning outcomes during CSL (Yu & Smith, 2011). Detailed analyses of eye movement data has also informed us of underlying cognitive processes in category learning studies (Rehder & Hoffman, 2005b, 2005a). This is because participants use their attention in the service of learning, by placing statistical biases on the learning data, thereby filtering information that enters internal processes. These filtering operations help to reduce uncertainty and shape how learning proceeds over time. However, apart from the studies cited above, few studies in statistical learning have focused on linking real-time selective attention with real-time information processing in statistical learning tasks. The goal of the present study is to look closely at the structure of these attentional biases during CSL.

Here we measure how visual attention is structured in time, and study how this organization may reflect underlying learning processes. We treat eye movements as a signal carrying information about participants' learning state, and model gaze patterns associated with prior knowledge. We study these gaze patterns using an unsupervised data mining method and present results that attempt to decode learners' internal knowledge from their eye movements. Our methods offer a general framework for measuring how selective attention and learning processes are coupled. But more importantly, they allow us to examine complex and highly variable eye movement dynamics at a fine grain, so that we may bet-

ter understand the cognitive processes which support cross-situational learning.

Eye Movement Data in Cross-Situational Learning

The gaze data in the present study were collected from a group of 61 adult learners. They were asked to learn 18 label-object mappings from 27 cross-situational learning trials. Each trial contained four objects displayed on a computer screen and four labels played sequentially in time. As shown in Figure 1, there were 4 perfect label-object mappings among 4 objects and 4 labels within a trial, but no information in the trial indicated which word went with which label. Participants used statistical evidence across multiple situations to build correct mappings, as shown in many previous experiments (Yu & Smith, 2011; Suanda et al., 2014; Yu et al., 2012). In this study, each trial lasted 11.25 seconds and started with a silent segment of 2.25 seconds before the first labeling event, allowing participants to become familiar with the set of 4 objects on screen before hearing the labels. The silent segment was followed by 4 labeling events with 2.5 seconds between label onsets. Labels were roughly 1 sec in length, so a temporal window of 2.25 seconds included both the labeling event itself and roughly 1.25 second after the labeling event as shown in Figure 1B. We recorded participants' eye movements using a Tobii 1750 eye tracking system over the course of 27 learning trials. At the end of the experiment, we tested participants' knowledge on each of 18 label-object mappings using an 18-alternative forced choice test. In addition, we also ran 3 other identical experiments, each with a unique set of subjects, the only difference being that either 3, 6 or 9 of the items had already been pretrained, so in those conditions we knew that learners already knew a specific subset of the 18 items. We refer to these conditions as 3pt, 6pt, and 9pt. Some results from this experiment were previously reported in Yu et al. (2012), where more details can be found.

In data preprocessing, we divided each learning trial into four temporal segments, one for each label as shown in Figure 1B. Over the course of 27 trials with 4 labeling events per trial, a total of 108 labeling segments were created for each subject. Within each segment we calculated the probability distribution for a subject's eye movements over the 4 on-screen objects. Here $P(o \mid l)$ is the proportion of time attending to object o given label l, reflecting the degree of gaze allocation to each object during a labeling event. As illustrated in Figure 1C, with 108 segments from each participant and a total of 61 participants for each condition, we calculated and collected 6588 (108*61) gaze distributions which we used as empirical data in the following analyses.

Using TCA to Examine the Structure of Sequential Visual Sampling Behavior

One way to study the structure of real-time selective attention is to use pre-defined behavioral patterns which describe certain aspects of how gaze is distributed (e.g. number of gaze

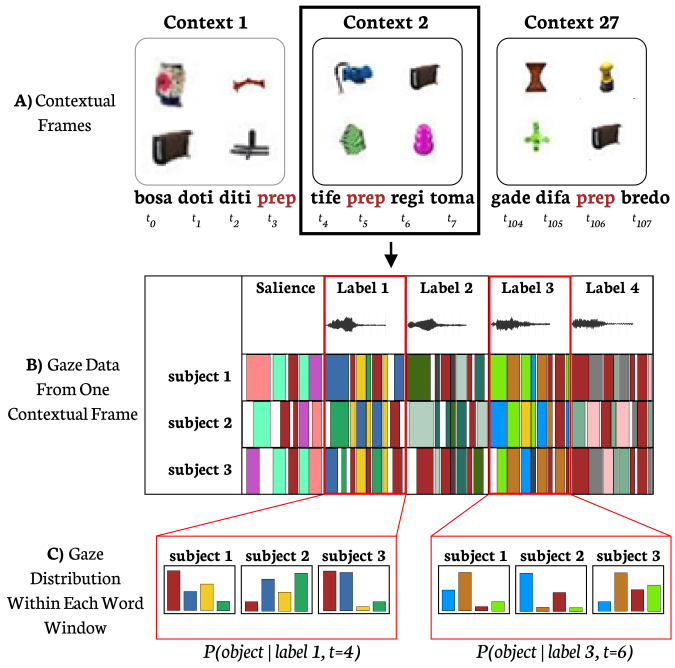

Figure 1: A schematic of the cross-situational learning experiment and its stimuli. In this figure, the target "*prep*" is a brown rectangular object, which subjects attend at varying degrees (red bars in panel B and C).

shifts, average fixation time, time to first fixation), and analyze how those behavioral patterns change over the course of learning. This approach can reveal important fine-grained behavioral differences across learners and has shown to be informative in explaining learning performance in CSL (Yu & Smith, 2011). By analyzing how gaze behavior shifts over the course of training, we're given a window into cognitive processes that support learning. However, given the complexity of natural eye movements, it's not *a priori* obvious which behavioral patterns we should choose in our analyses and why these particular patterns are useful for understanding underlying cognitive processes. The present study proposes an alternative approach in eye movement analysis – learning these patterns from the gaze data itself and modeling their covariation in time.

In order to infer useful gaze patterns from raw data, we used an unsupervised factor analysis method – the canonical polyadic tensor decomposition (CPD) (Hitchcock, 1927). CPD is a multilinear method for dimensionality reduction, modeling tensor-structured input as a low dimensional linear projection. These projections form the basis of how we describe attention behavior. What's unique about CPD is that it allows us to model how attention is organized at multiple time scales, both within and across trials, by treating different scales as separate factors encoded along the different axes of an input tensor[1]. In a recent paper by Williams et al. (2018),

[1]Tensors are a generalization of vectors and matrices, where vec-

CPD has been used to successfully decode functionally specific neural activities during learning tasks, and went by the name of Tensor Component Analysis (TCA). This is the name we'll use for it here.

TCA is similar to Principal Component Analysis (PCA) and other matrix factorization techniques, but offers a few key advantages. With TCA, we can model how multiple factors interact (e.g. subjects, within-trial gaze, and across-trial gaze) by decomposing gaze data into linearly independent components, finding *unique* representations that unmix their interactions to best explain the gaze data. In PCA, similar analyses require unrolling inherently tensor-shaped data into a 2D matrix representation, as a result destroying the organization present in its equivalent tensor form. Compared with TCA, PCA's decomposition introduces what's known as the "rotation problem", where any given factorization specifies an infinite number of solutions (under arbitrary rotations). Rotations keep reconstruction loss fixed while transforming how these solutions are organized. In order to tie structural aspects of attention to cognitive processes, we need to know specifically how these solutions are structured, which presents a significant problem. To solve this, PCA constrains solutions to be orthogonal and orders its components in terms of the amount of variance explained. However, these constraints are overly strict, as we have no reason to assume that attention is composed of orthogonal components. We leverage TCA's uniqueness properties (Kruskal, 1977) to model within- and across-trial variation in subjects' gaze behavior, avoiding this specific shortcoming of PCA. We encode subjects' gaze distribution dynamics as a 3rd order tensor, and use TCA embeddings to model attention dynamics during CSL.

Each axis of an input tensor $\mathcal{X}$ maps to a separate factor after decomposition, with each factor represented as a 1D vector. When factors are recombined, they reconstruct the original behavioral data. As shown in Figure 2, $\{\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r\}$ are the learned factors (one for each of the 3 axes of $\mathcal{X}$), "$\circ$" is the vector outer product, and R is a hyperparameter controlling the number of components that TCA learns. We organized subjects' gaze distributions into a 3rd-order tensor, so that axes partition data along three meaningful dimensions – subjects, gaze distribution at label onset, and gaze distribution across the 108 labeling events. We used these factors to model similarity in subjects' otherwise disparate selective attention dynamics, reducing complex sequential behavior to fixed points in an embedding space (one point for each subject). TCA places minimal top-down constraints on the learned representations, only stipulating that components are linearly independent. In our case, we additionally constrained solutions to be non-negative.

To illustrate how TCA decomposes gaze behaviors, we present a toy example with synthetic gaze data in Figure 3. We randomly generate gaze distributions for simulated learners, where their attention to the 4 on-screen items is randomly

tors are 1st order tensors and matrices are 2nd order tensors. Our analyses structure behavioral data as a 3rd order tensor. See Kolda and Bader (2009) for a review of tensor decomposition.

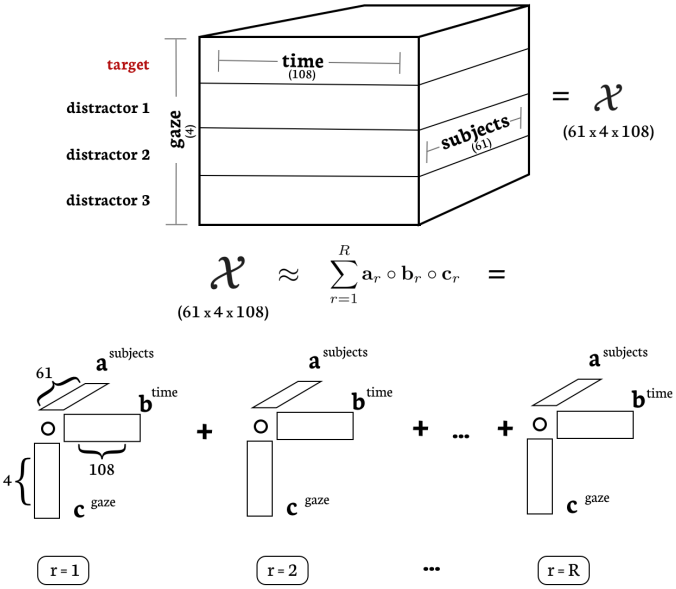

Figure 2: Subjects' gaze is organized as a 3rd order tensor and represented as the sum of R outer products using TCA. These are the R "components" of attention behavior. For the gaze axis, we encode the proportion of time spent looking at the 4 items on screen after a label onset. The first row in the gaze axis codes for proportion of time looking at the target, with subsequent rows coding for proportion looking to distractor (in decreasing order, where distractor 1 is most attended, followed by distractor 2, then distractor 3). Gaze distributions are arranged as 108 column vectors along the time axis, corresponding to 108 label-onset instances in the experiment. Sampling trajectories for each subject are slices arranged along the third axis, one for each of the 61 subjects.

distributed as a Gaussian whose mean is shifting over time ($\tilde{p}(x,t) = \mathcal{N}(\mu_t, 1)$, where μ_t is a function of time). These synthesized gaze data form the column vectors running along the time axis in $\mathcal{X}$, with each simulated learner as a slice along the subjects axis. We randomly partition subjects into two groups, A and B, and give them two distinct time dynamics: in Group A, their μ_t shifts from 0 to 4 across the 108 time points (in equal increments), and in Group B this shift goes in reverse, with μ_t going from 4 to 0. TCA differentiates these two kinds of subjects by recovering the underlying dynamics of their gaze, represented as a mixture of 3 linearly independent components (one for each value of R). Each of the 61 simulated subjects are indexed along the x-axis of the subjects factor. Subjects in Group A are highlighted in yellow. In the r_3 component we see that Group A subjects over-weight the "low-μ decreasing over time" gaze pattern. In the r_1 component, the group B subjects show the inverse pattern, while the r_2 component does not code for any cross-subject variation. As shown in Figure 3, TCA successfully decodes the three latent variables as subject factor, gaze factor and time factor, to perfectly match the raw synthesized data. Thus, this toy ex-

ample shows how TCA uncovers key factors from observed data.

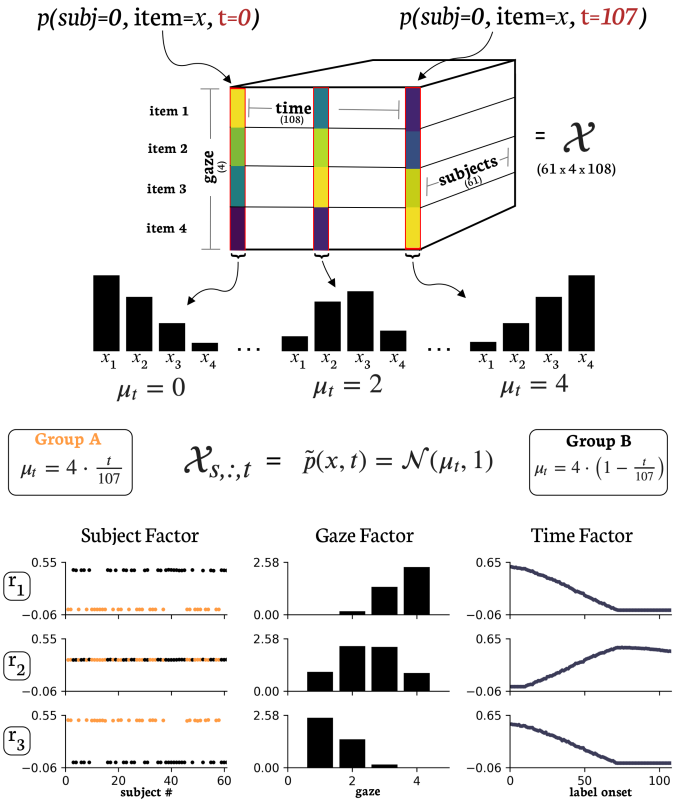

Figure 3: We compute an $R = 3$ decomposition on synthetic gaze distributions and plot the resulting factors.

Study 1: Eye Movements Predict Learning Dynamics

We applied TCA at two levels, one at the *subject level*, and the other at the *item level*. Figure 2 shows a schematic of the subject level decomposition, where individuals' behavior is encoded along the axis dedicated to subjects. For the item level decompositions, we collapse across subjects and encode individual item gaze distributions in place of the subjects axis. Instead of 108 entries for the time axis, there are only 6 time points in the item-level analyse since each item is labelled 6 times across an experimental condition when paired with other distractor items. These two separate decompositions serve to model variability in both subject level attention, as well as attention specific to individual items.

Our analyses present decompositions where $R = 3$, as we find this setting leads to a high degree of similarity in solutions across many optimization runs, while significantly minimizing reconstruction loss. See Figure 4 for plots of subject level decomposition weights. Here TCA uncovers 3 distinct components of gaze behavior: $\mathbf{r_1}$: selection that's spread across both target and distractor items and decreasing over time, $\mathbf{r_2}$: skewed to the distractors and increasing over time, and $\mathbf{r_3}$: skewed to the target and increasing over time. Along with the gaze factors, TCA also learns 2 other factors – subject and time. These correspond to the other 2 axes of $\mathcal{X}$.

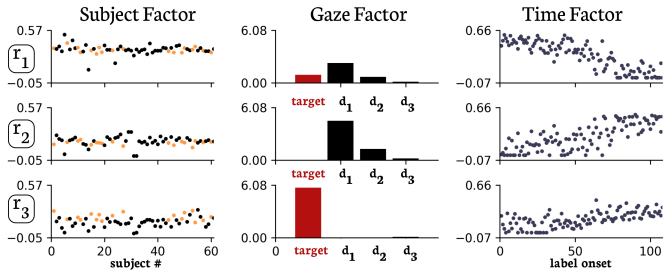

Figure 4: In a subject-level decomposition (with $R = 3$), TCA uncovers how 3 unique types of gaze dynamics should be mixed to reconstruct the input. We took a median split on subjects as a function of the number of items they learned at the end of the experiment, with strong learners (≥ 6 items) highlighted in orange.

Attention Space The factor weights indicate the degree to which other factors in that component mix to reconstruct $\mathcal{X}$. This means we can use the subject and item factor weights to define an R dimensional "attention space", wherein each subject (or each item), is a point corresponding to the different values encoded by their R distinct subject (or item) factors. See Figure 5.

This reduces subjects' highly variable sequential sampling dynamics into a common frame of reference, thereby allowing us to model the relationship between visual selection and learning. We took a median split to divide subjects into a strong learner group and a weak learner group based on the number of items learned by individual subjects. If strong learners tend to occupy a certain behavioral subspace, this would suggest a commonality in how they deployed their attention which ultimately led to successful learning. If strong and weak learners were not separable in the behavioral space, it would suggest we cannot use gaze behavior to predict learning outcomes. In the following subsections, we will present a classification approach to examining whether we can use TCA-based representations to separate strong from weak learners in the subject space, and to predict learned or unlearned items in the item space.

Using TCA-based Representations to Predict Learning

We used classification accuracy via linear kernel Support Vector Machine (SVM) to measure how well gaze behavior can be used to separate learning outcomes at both the subject and item levels. We treated classification tests as a proxy measure for the amount of information gaze carries about internal learning states. The classification results we report include mean accuracies in nested cross validation runs of non-negative CP decomposition (n=100, leaving random 10% of subjects/items out) and classification on held out subject/item

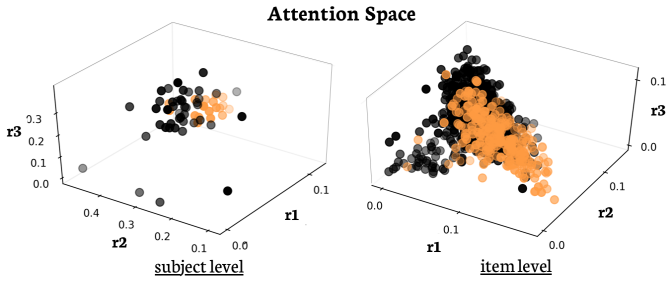

Figure 5: Attention spaces for subject- and item-level TCA decomposition. In the subject level space, strong learners are highlighted in orange. For the item level space, orange points indicate items that were learned by subjects at the end of training.

embeddings (n=100, random 20% left out) using linear SVM. Decompositions were fit using Hierarchical Alternating Least Squares (HALS) method, with a stopping tolerance of $1e{-}5$ for reconstruction error. For subject and item-level learning classifications we omit the 9pt condition as learning performance was at ceiling.

Predicting Subject-Level Learning Strong learners (by median split on the number of learned items) occupy a unique neighborhood in the behavior space, allowing us to correctly differentiate between them in all of the three experimental conditions (0pt: Acc=0.71, SD=0.06; 3pt: Acc=0.76, SD=0.06; 6pt: Acc=0.74, SD=0.06).

Predicting Item-Level Learning We were also able to predict learning outcomes for individual label-object mappings in all experimental conditions (0pt: Acc=0.76, SD=0.01; 3pt: Acc=0.62, SD=0.01; 6pt: Acc=0.77, SD=0.01).

Decoding Learning States From Eye Movements In pretrained item conditions, already-known items are mixed alongside to-be-learned items during training. Stimuli and their presentation order are identical in all 4 conditions, the only difference being subjects' relative degrees of prior knowledge about label-object mappings. This means we can compare attention behavior for *specific items* when a) they're already known, relative to b) when those same items are not known.

We run three analyses (3vs0, 6vs0, and 9vs0), each testing discrimination performance on pretrained items versus their identical non-pretrained counterparts taken from the 0pt condition. We find 3vs0 classification is at chance (Acc=0.52, SD=0.01), meaning we're unable to detect whether a learner already knows a specific item. However, in both the 6vs0 and 9vs0 classifications we're able to successfully decode item knowledge from subjects' gaze patterns above chance (6vs0: Acc=0.62, SD=0.06; 9vs0: Acc=0.64, SD=0.06), suggesting that while partial information in the 3pt condition may still be used internally (e.g. mutual exclusivity judgements), there may be a threshold in the amount of prior item knowledge necessary to influence the structure of visual selection.

Associative Tensors In the previous decomposition, the gaze axis of our input tensors coded distractor items in decreasing order of looking time (determined within a single label onset window). See Figure 2. By structuring the tensor in this way, we include no information about how those specific items had co-occurred with the target. We build a separate set of tensors which encode this information. Instead of treating distractors as blank slates at each label onset, we track label-object associative strengths, proportional to subjects' past accumulated sampling for those label-object pairs, ordering these associates along the gaze axis in decreasing associative order. These tensors allow higher decoding performance for pretrained items compared to their blank-slate-distractor counterparts (6vs0: Acc=0.65, SD=0.05; 9vs0: Acc=0.68, SD=0.02; significant by Wilcoxon test, $p < 0.001$). As in the previous analyses, we're still unable to decode 3vs0 pretrained items above chance using associative tensors (Acc=0.52, SD=0.04).

Study 2: Fine-Grained Behavioral Analysis

Our analyses in Study 1 demonstrate how item knowledge leaves a unique signature on eye movements, allowing us to differentiate 3 classes of items: learned, not-learned, and pretrained. In Study 2, we look closer at what this signature looks like. Since TCA factors are interpretable (and unique) representations of the original input, we study how these solutions might represent underlying cognitive processes, looking closer at how behavior differs in these 3 types of items. See Figure 6 for item-level decompositions across different pretrained conditions.

We find a number of distinguishing patterns across the 3 item categories. For the learned items, in all pretraining conditions, components coding for target looks show smooth increases in gain over time, with corresponding smooth decreases in the distractor look components. On the other hand, in not-learned items, looks to target demonstrate a more uneven pattern. In the 3pt and 6pt conditions, not-learned items show sudden drops in target looking at the second and fifth label onsets. Otherwise, their patterns are flat (early peaks are as large as middle and late peaks). Like the learned items, the not-learned items show similar decreases in distractor looking over time. In the 0pt condition, attention for not-learned items is dominated by distractor looking throughout training. Pretrained items show similar general trends as the learned items, namely gain in target looking and decrease in distractor. However, pretrained items distinguish themselves at 2 time points. In the 3pt condition, there are 2 drops in target looking, one at the 3rd label onset and the other at the 6th label onset. In the 6pt condition, these occur at the 4th and 6th labeling events. These drops are matched with mirroring increases in distractor looking at these same time points. These patterns likely reflect prior knowledge drawing attention away from useless information (i.e. the target, as they already know it), directing it towards to-be-learned items.

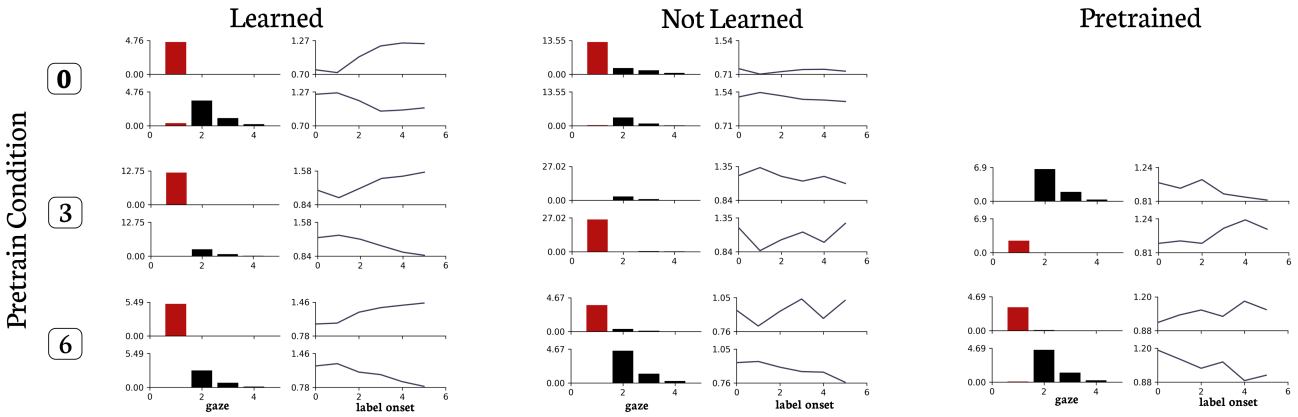

Figure 6: Item and time factors for a 3-way partition of items - 1) learned, 2) not learned, and 3) pretrained. Plots reflect independent $R = 2$ decompositions, where input tensors contain items of a single type.

Discussion

Eye movements offer a rich behavioral signal which can be used to extract detailed information about internal states of knowledge. However, in most eyetracking studies only a small proportion of the total variance is used. Here we've applied TCA in order to make better use of the rich variation found in natural gaze behavior, going beyond simple looking measures as an index of learning. We link complex gaze patterns to subjects' internal knowledge as they learn label-object pairs through statistical learning. Our methods offer a general framework for extracting behavioral markers of learning, opening the door for future studies that use these markers as part of their experimental designs. For example, experimental stimuli might be presented contingently as a function of subjects' real-time knowledge states, as inferred through their eye movements.

In future work we hope to extend our analyses from Study 2. Prior studies have shown that sustained, and stable, attention in CSL is a strong predictor of learning at the subject level (Yu & Smith, 2011; Yu et al., 2012). The gaze patterns we find in not-learned items may reflect sequential aspects of instability at the *item level*, as evidenced by their lack of smooth gain in sustained attention over time, as shown in Figure 6. Because TCA models sequential similarity in gaze dynamics, we can track how sustained attention is organized *across* labeling events for individual items. In contrast, measures of sustained attention that focus within a single label onset window, or that take averages across labeling events, are unable to resolve this level of fine-grained organization in attention behavior. Our results suggest that, at the item level, failure to learn may be associated with a specific form of low sustained attention – one that's fluctuating in the course of learning with occasional sudden drops in sustained target looking. In future work, we will design experiments to explicitly test this hypothesis. In general, data mining and explicit hypothesis testing can go hand in hand to advance our understanding of the learning mechanisms underlying cross-situational learning. Data mining techniques allow us to discover fine-grained behavioral patterns that we can form specific hypotheses about and design well-controlled experiments to test.

Another contribution of the present study was the application of TCA in the analysis of fine-grained behaviors. Even though similar embedding techniques have shown promise in modeling complex sequential behavior (Dezfouli et al., 2019), we suggest that our presented approach offers a unique advantage when it comes to interpretability. The main strength of TCA doesn't lie in its absolute decoding performance (other models will likely outperform it), but in the transparent meaning of its decomposed factors. With TCA, behavior and latent representations are related by a simple linear map, where both behavior and latent spaces share a common semantics in their organization. In contrast, while it's possible to interpret embeddings from more complex non-linear estimators (e.g. Variational Autoencoders and LSTMs with various disentanglement techniques), linking the organization of their solutions to the original behavioral processes is less straightforward. Because of these unique properties, we can systematically introduce top down constraints on TCA's solutions by restructuring inputs to address specific research questions, tracking precisely how these manipulations influence the organization of the latent space. Our analyses with associative tensors are a first step in that direction. In addition to gaze, TCA can be applied to any type of high-density data to extract uniquely predictive spatiotemporal patterns, and to use these to infer latent variables and structures in cognitive processes.

Acknowledgments

We thank Linda Smith, the CCL Lab, and Eric Bigelow for helpful discussions during the preparation of this paper. This work was funded by National Institute of Child Health and Human Development R01HD074601 and R01HD093792

References

Dezfouli, A., Ashtiani, H., Ghattas, O., Nock, R., Dayan, P.,

& Ong, C. S. (2019). Disentangled behavioural representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 2254–2263). Curran Associates, Inc.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a zipfian environment. *Cognition*, *189*, 11–22.

Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, *6*(1-4), 164–189.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455–500.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, *18*(2), 95–138.

Quine, W. V. O. (1960). *Word and object*. MIT press.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*(1), 1–41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of experimental child psychology*, *126*, 395–411.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., … Ganguli, S. (2018). Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, *98*(6), 1099–1115.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180.

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective attention in cross-situational statistical learning: evidence from eye tracking. *Frontiers in psychology*, *3*, 148.