

Visual Knowledge Tracing

Neehar Kondapaneni¹ Pietro Perona¹ Oisin Mac Aodha²

¹Caltech ²University of Edinburgh

Abstract. Each year, thousands of people learn new visual categorization tasks – radiologists learn to recognize tumors, birdwatchers learn to distinguish similar species, and crowd workers learn how to annotate valuable data for applications like autonomous driving. As humans learn, their brain updates the visual features it extracts and attend to, which ultimately informs their final classification decisions. In this work, we propose a novel task of tracing the evolving classification behavior of human learners as they engage in challenging visual classification tasks. We propose models that jointly extract the visual features used by learners as well as predicting the classification functions they utilize. We collect three challenging new datasets from real human learners in order to evaluate the performance of different visual knowledge tracing methods. Our results show that our recurrent models are able to predict the classification behavior of human learners on three challenging medical image and species identification tasks.

Keywords: visual classification, knowledge tracing, human learning

1 Introduction

Humans excel at learning new concepts even when they have only received limited explicit supervision [28,55]. Key to our success is our ability to extract informative and generalizable representations from the world around us and our ability to update these representations given relatively sparse feedback. This capacity, in turn, enables us to perform complex tasks such as spatial navigation and visual categorization with apparent ease.

Despite recent progress that has been made in computer vision in learning visual representations through self-supervision alone [54,9,14], large amounts of supervision are still required to make best use of the resulting features [12]. In light of this, it is important for us to better understand: (i) what are the properties that make representations learned by *humans* so effective, (ii) how are these representations learned, and (iii) can we predict the classification behavior of humans during learning? Our ultimate goal is to obtain better insight into how humans are such effective learners, which can then potentially inform new learning mechanisms for future artificial systems.

In order to attempt to address some of these questions, in this work we explore the problem of visual knowledge tracing. In the educational data mining community, knowledge tracing is the problem of monitoring and predicting the

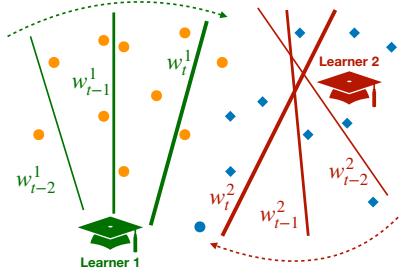


Fig. 1: We model a learner as an evolving classifier in a learned feature space. We assume the feature space is static and, as the learner is presented with images and class labels over time, their internal classification function self-updates. Here, we illustrate this for three time-steps, for two learners, learning a binary classification task (orange versus blue).

evolving knowledge state of a learner engaged in a learning task [26]. Recent work has applied advances in deep learning to knowledge tracing for question and answer-style text datasets and has investigated applications in domains such as mathematics education [34,33,35]. However, tracing the behavior of humans that are engaged in learning and performing challenging visual categorization tasks is underexplored. Most closely related is the work on deep metric learning that attempts to learn human aligned visual representations from sparse human annotations [19]. However, these works typically make simplifying assumptions, e.g. assuming the learners are not changing over time (i.e. they are ‘static’) or that the visual criteria used by the learners is the same across all learners. Instead, we explore a more challenging visual knowledge tracing setting where the learners are assumed to be non-stationary during learning, i.e. the visual features they use to perform the classification task at hand can, and likely do, change over time (see Fig. 1).

We present a recurrent neural network-based approach for visual knowledge tracing. Once trained, our models are capable of predicting the classification behavior of human learners that were not observed during training. The proposed models make use of the history of previous learner responses, images, and ground truth class labels in order to predict their future responses. Through experiments on three challenging image classification datasets we show that our models are superior to baseline approaches. Our models are capable of tracing the learning dynamics more accurately than non-recurrent baselines.

We make the following three contributions: (i) A new model for visual knowledge tracing that jointly estimates the visual features and per-time-step classification function used by non-stationary human learners. (ii) A new set of annotations for three benchmark evaluation datasets collected from humans engaged in learning challenging visual classification tasks. (iii) A detailed comparison of several visual knowledge tracing methods on these datasets.

2 Related Work

2.1 Metric Learning

The goal of metric learning is to learn perceptual embeddings such that distance in the lower dimensional embedding space encodes information related to semantic similarity. Pre-deep learning approaches to metric learning were primarily concerned with learning embeddings directly for each item in an input set. In the case of learning from human supervision, approaches that use relative similarity judgements have been shown to be effective [23,45,53,37]. These

methods have also been extended to the adaptive setting where the model can decide which items to request annotations for in order to speed up training [43].

More recently, end-to-end metric learning methods have attempted to parameterize an embedding function (e.g. a convolutional neural network) directly [38,29]. As a result, they are able to embed any new item into the embedding space, even those not seen at training time. Representative earlier applications of this line of work include image ranking [49] and face recognition [40]. The standard assumption made by the majority of these methods is that only *one* similarity criterion is being used. However, when collecting data from human annotators, different individuals may be using different visual criteria when making classification and similarity judgements, e.g. one individual could be using shape, while the other is using color. Furthermore, the same individual may change the criteria they use conditioned on the specific items they are shown, and could change to another criteria when shown another set of items at a different point in time.

There has been some work that attempts to deal with the fact that *different* similarity criteria may be being used. These range from fully supervised, where the similarity criteria is known at training time [48], through to unsupervised methods that attempt to estimate the criteria [21,30,44,52]. Attempts have also been made to probabilistically estimate item embeddings along with annotator-specific parameters representing the individual criteria they are using [52].

One common assumption made by the above methods is that the learner is stationary, i.e. they are used a predefined and fixed similarity criteria, or small set of criteria, which do not evolve or change over time. This is a reasonable assumption to make when dealing with common everyday object categories where the annotators will likely be familiar with the objects depicted and have an a priori understanding of how the visual features of the objects may vary. However, this assumption is violated in cases where the annotator is in the process of *learning* the visual concepts of interest. In this work, we address this non-stationary setting and show that by doing this, we can more accurately predict the visual classification behavior of real human learners.

2.2 Human Category Representation

Existing models of human category representation and learning can be clustered into five major groups: rule-based, prototype-based, exemplar-based, knowledge / theory-based, and decision boundary-based approaches [56,4]. The current consensus is that humans likely use multiple different category-learning systems depending on the specific nature of the task at hand [5,3]. For example, in rule-based tasks, the optimal policy may be easy to verbalize and thus efficiently encoded via a set of rules. In practice however, perceptual tasks such as fine-grained visual categorization can be much harder to represent in this way [8]. Several works have attempted to extract perceptual embeddings that align with human similarity judgements from coarse [15,36] and more fine-grained [32] image collections. [6] showed that with simple linear transformations, pre-trained deep image classifiers can be predictive of human similarity judgements. Relevant

to our work, [7] investigated whether human learning dynamics mimic gradient descent in Artificial Neural Networks when learning visual categories. However, they assumed they had access to the feature space used by the human, whereas we instead attempt to learn this. In this work, we also aim to extract human-aligned representations, but in the more challenging setting whereby our learners are not static, but instead are in the process of learning the categorization task.

2.3 Knowledge Tracing

The problem of modelling the hidden state of dynamic learners as they interact with a learning task has also been tackled in the knowledge tracing literature. Bayesian Knowledge Tracing-based methods model a learner’s knowledge state by assuming that the learner can be represented as a Markov process which updates online during learning [13]. Building on this line of work, Deep Knowledge Tracing (DKT) instead uses a recurrent neural network as the underlying tracing model [34], and fully self-attention-based methods have also been proposed [33,35]. It is important to note that conventional knowledge tracing attempts to model knowledge acquisition as a binary variable at the ‘skill’ level (i.e. the visual class) as opposed to the ‘instance’ level (i.e. a specific image). In contrast, our approach jointly learns an image embedding function in addition to being able to capture and predict the individual learning trajectories of multiple different learners. A detailed comparison of the original DKT model and our model is provided in the supplementary material (see Sec. A.2).

2.4 Machine Teaching

Estimating the representations used by humans is an important component for developing automated teaching algorithms and systems. Machine teaching algorithms address the teaching problem by generating sequences of instructional examples to show to novice learners in order to improve their ability on a given task [57]. Machine teaching has applications in crowdsourcing where the aim is to efficiently train crowdworkers, in addition to education where the goal is to train new experts, e.g. in medical image analysis [10,2].

There is a growing body of work in computer vision that attempts to teach visual concepts to human learners, e.g. [42,18,27,50,51]. However, many of these works assume a fixed feature space that is generated before teaching begins [42,18,27]. In one experiment, [27] showed that representations that are better aligned with human perception result in improved learner performance on the downstream teaching task. While we do not explicitly investigate teaching algorithms in this work, we instead explore a setting where data is collected from humans engaged in learning a visual categorization task with instructional images selected by a ‘random’ teacher. Importantly, the representations and learner parameterizations extracted by our model can be used directly with computer assisted teaching methods.

3 Method

Our goal is to estimate the image classification function used by a human learner that has been provided with a sequence of images and corresponding ground truth class labels as training data. We begin by outlining the problem, and then present our approach to human visual knowledge tracing.

3.1 Problem Setup

Given an image \mathbf{x} as input, we model a human learner as a classification function that returns a response, $r^k = \text{argmax}_c P(c|\mathbf{x}, \theta^k)$. Here, r is a discrete class label representing the class response for learner k , i.e. $r \in \{1, \dots, C\}$, where C is the number of possible classes, and θ^k are unobserved parameters representing the state of the learner. Our learners are not stationary as their internal ‘knowledge state’ changes depending on the information they have previously been exposed to that is relevant to the task. As a result, for a given learner k we model their classification function at time t as $r_{t+1}^k = \text{argmax}_c P(c|\mathbf{x}, \theta_t^k, \mathbf{x}_{1:t}^k, y_{1:t}^k, r_{1:t}^k)$. Here, $(\mathbf{x}_{1:t}^k, y_{1:t}^k, r_{1:t}^k)$ is the history of images, ground truth class labels, and responses that a learner k has seen, and provided, up to and including time t .

Specifically, at each training time-step, a learner k is presented with an image \mathbf{x} , they provide their response r , and are given feedback in the form of the correct class label y (Fig. 2B). However, fitting a model for an individual learner with a single response per time-step is difficult. Alternatively, requesting more responses per time-step would reduce the number of teaching examples presented to the learner in the same amount of time. Instead, to overcome this limited information setting, we train a model ϕ across many learners, allowing the model to discover knowledge states and learning rules shared across all learners. Once trained, our model can make predictions for how a learner, who was not observed during training, will classify an image based on their prior classification behavior.

3.2 Tracing Human Learners

Our tracing model ϕ can be decomposed into a feature extractor f , a classification function ψ , and a non-learned and non-linear transformation σ (softmax). We explore how to represent the feature extractor and classification function.

A natural choice for the feature extractor f is a Convolutional Neural Network (CNN). Given an image \mathbf{x} as input, the feature extractor outputs a D dimensional vector $\mathbf{z} = f(\mathbf{x})$. We will assume that all learners use the same underlying feature extractor which remains constant over time i.e. $f = f_t^k$, and that they simply differ in the relative importance they place on different visual features. While these are both big assumptions to make, they are not overly restrictive. For example, a novice and an expert might engage in the same visual classification task but differ in the set of visual features they select in order to make their decision. Furthermore, while we assume that the feature extractor remains constant over the time interval of our experiments, we do not assume that the classification function ψ used by a learner remains static. In the next sections

we will explore different choices for this classification function, comparing simple static classifiers with more expressive recurrent models.

Static Tracing Model. The first model we explore is the simplest. Here we assume that all learners use the same classifier which does not vary over time. In this setting, ψ is a multi-class linear classifier with a weight matrix \mathbf{w} and per-class biases \mathbf{b} ,

$$\phi_{static}(\mathbf{x}) = \sigma(\psi(f(\mathbf{x}))) = \sigma(\mathbf{w}^\top f(\mathbf{x}) + \mathbf{b}). \quad (1)$$

This model is similar to conventional metric learning approaches which do not attempt to capture any annotator specific differences related to individual biases or temporal changes. At training time we simply estimate one set of parameters for all learners. This model does not take the response history into account.

Time-Sensitive Tracing Model. One obvious limitation of the static tracing model is that it does not take into account the fact that a learner will likely change over time, i.e. they may be much worse at a new classification task early on, but may improve over time as they are shown sequences of example images along with their associated ground truth class labels. A more advanced model, that captures this temporal evolution, is one that has a different classifier for each time-step,

$$\phi_{static_time}(\mathbf{x}) = \sigma(\mathbf{w}_t^\top f(\mathbf{x}) + \mathbf{b}_t). \quad (2)$$

Again, the same classifiers are shared across all learners, but in this case the weights and biases are different at each time-step, i.e. $\mathbf{w}_t \neq \mathbf{w}_{t-1}$.

3.3 Recurrent Tracing Models

The previous tracing models do not account for the fact that individual learners may start with different levels of ability and update their internal knowledge state in different ways depending on the information that they are provided with. [34] showed that recurrent networks could be used to track the skill acquisition of human learners engaged in learning math quiz questions. Direct application of their model to our visual categorization setting is not possible as they assume one hot encodings of the query and learner responses as input. Their approach also uses large training sets – on the order of thousands of learners and tens of thousands of interactions. Furthermore, they model knowledge acquisition at the ‘concept’ (i.e. visual category) and not ‘instance’ (i.e. a specific image) level, and thus their approach is not capable of making predictions for items not seen during training. We build on [34] and adapt it to our visual category learning setting by presenting two different recurrent-based models for human visual knowledge tracing.

Direct Response Model. Our first model uses a recurrent network to directly predict the responses of a learner given their previous response history,

$$\phi_{direct}(\mathbf{x}) = \sigma(\psi_{rnn}(\mathbf{z}_{1:t}^k, y_{1:t}^k, r_{1:t}^k, \mathbf{z}, y)). \quad (3)$$

Here, ψ_{rnn} is a recurrent network (in practice we represent this using an LSTM [16]) and $\mathbf{z}_t = f(\mathbf{x}_t)$ are visual features extracted from our CNN.

This model assumes that a learner’s knowledge state at time t is defined by the images they have previously seen and their past classification responses. Recurrent models can produce unique transformations for individual learners by conditioning on their hidden states. In this case, after the shared feature extractor transforms an image into a feature vector, the model modifies the feature vector with a series of non-linear transformations conditioned on the learner’s hidden state. The final linear layer transforms the feature vector into a predicted response. Note that this model is also conditioned on the current query image $\mathbf{z} = f(\mathbf{x})$ and the corresponding ground truth class label y .

Classifier Prediction Model. Our second recurrent model attempts to provide a more interpretable approximation of human classification. In this case, instead of letting the recurrent model directly predict the probability for each response, it instead attempts to approximate the weights of a linear classifier used internally by the learner. Importantly, the values this classifier takes will depend on the response history of a given learner and will differ at each time-step,

$$\mathbf{w}_t^k, \mathbf{b}_t^k = \psi_{rnn}(\mathbf{z}_{1:t}^k, y_{1:t}^k, r_{1:t}^k, y), \quad (4)$$

$$\phi_{cls_pred}(\mathbf{x}) = \sigma(\mathbf{w}_t^k \mathbf{z} + \mathbf{b}_t). \quad (5)$$

Unlike the previous direct prediction recurrent model, we now explicitly represent the classification function used by an individual learner. Also, here the features \mathbf{z} for the query image are not processed by the recurrent network in Eqn. 4. Instead they are evaluated using the much simpler predicted classifier weights in Eqn. 5. This decoupling is advantageous in applications like machine teaching where we have to query the tracing model multiple times at each time-step in order to determine the next image to show learners. Reducing the computation required to perform these queries will result in faster teacher algorithms.

3.4 Training Tracing Models

We jointly estimate the parameters of the feature extractor f and classification functions ϕ for each of the above models using a standard cross entropy loss,

$$\mathcal{L} = -\frac{1}{KT} \sum_{k=1}^K \sum_{t=1}^T \log(\phi(\mathbf{x}_t^k)_r). \quad (6)$$

Here, $\phi(\mathbf{x}_t^k)_r$ indicates the predicted probability from a model ϕ choosing class r , for learner k , at time-step t . The training objective aims to minimize the

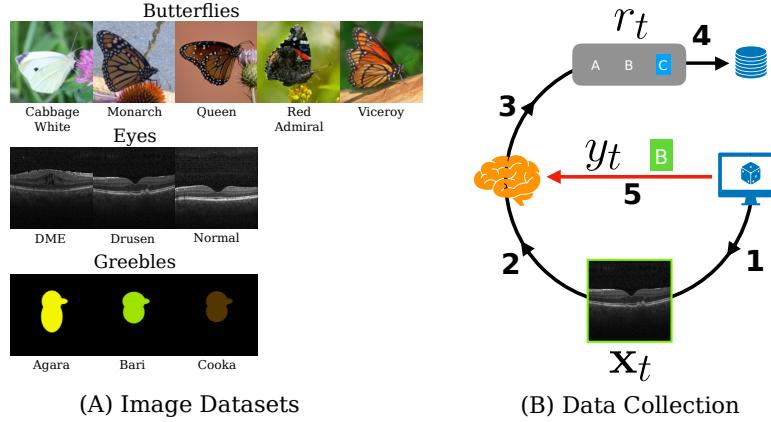


Fig. 2: **(A)** Example images from the three different datasets from our experiments. ‘Butterflies’ contains images of five different species and was originally presented in [27]. ‘Eyes’ contains optical coherence tomography images of the human retina from [20], and features two diseased classes and one normal one. ‘Greebles’ is a synthetic dataset we created where the three object classes vary in terms of shape and color. **(B)** Data collection pipeline. A random image is selected (1), shown to the learner (2), and the learner provides a response (3). Their response is stored (4) and the correct class label is provided to them (5).

difference between the learner responses from our training set and the tracing model outputs.

4 Experiments

In this section we evaluate the different proposed models for visual knowledge tracing on data we collected from real human participants¹.

4.1 Datasets

Traditional image classification datasets mostly contain labels produced by annotators familiar with the subject material, e.g. [25,46,24], or they have at least received detailed instructions and examples on how to annotate them, e.g. [39]. As a result, these datasets do not contain annotations from learners engaged in learning a task and are thus not suitable for evaluating visual knowledge tracing. While some work has focused on teaching crowd learners (e.g. [42,27,50]), they often use teaching image sequences that are determined offline and fixed. For our tracing experiments, we require unbiased sequences of images that are randomly selected for each learner. Some of these existing works compare their approaches to a random image selection baseline, but the size of these random

¹ Code and dataset - <https://github.com/nkondapa/VisualKnowledgeTracing>

teaching subsets is insufficient for thorough evaluation of our different tracing approaches, e.g. [27] have random selection data from only ~ 40 participants. Due to these limitations, we collected annotations from human learners for three challenging fine-grained visual classification datasets.

Image Data. We selected three different image datasets that cover three distinct domains: artificial data where we have full knowledge of the underlying distribution, medical image data, and images of different wildlife species. The first two datasets in particular are representative of the types of visual identification tasks that many humans are interested in learning.

Our first dataset, ‘Butterflies’, contains images from five different common species of North American butterflies. The ‘Cabbage White’ class is immediately recognizable, ‘Red Admiral’ can be learned relatively easily, and the remaining three are difficult to discriminate. This dataset was originally used in [27] and contains between 386 and 481 images per class, for a total of 2,224 images. Our second dataset, ‘Eyes’, is a three-class subset of a large collection of publicly available images of the human retina from [20]. It contains two diseased classes, ‘Diabetic Macular Edema’ (DME) and ‘Drusen’, and one ‘Normal’ class. We manually selected 200 images from each class. The third dataset is a challenging synthetic one we created called ‘Greebles’. It contains three classes, where the underlying feature space used to generate the images is known by design. The relevant features are the body length and color, but the images also includes some irrelevant variation in the form of the head size and body width. The distinctions between the classes can be subtle, making the task challenging. It contains 1,200 images in total, with an even number per class. Visual examples for each of the datasets are presented in Fig. 2A. Note that the single examples in Fig. 2A do not covey the visual diversity of the datasets.

Human Data Collection. For each of the previously described datasets, we collected data from human participants that were engaged in learning the classification task. Each learner was presented with 30 training images and 15 test images. During the training phase they were provided with ground truth feedback indicating the correct class labels. This feedback was not provided in the test phase. For each image, we asked learners to rank the top three classes in order of most likely to least likely to be correct (see Sec. F.1 for further information). The training and testing examples were randomly selected for each learner. An overview of the data collection process for one iteration, for one learner, is shown in Fig. 2B.

The data was collected using a custom built a web application and the participants were recruited through the crowd sourcing app Prolific [1]. In total, we collected data from 150 learners for each dataset, where each individual could only do the task once. The median time spent on the Butterflies task, including training and testing, was ~ 11.7 minutes with a median of 11 correct on the test phase. The corresponding statistics for the Eyes dataset was ~ 12.1 minutes

with a median of 13 correct, and for the Greebles dataset is was ~ 8.8 minutes with a median of 9 correct.

Our human learners demonstrated learning across all three datasets. In Fig. 3 we plot two histograms for each dataset, the first histogram displays the number of correct responses during the training (i.e. teaching) phase and the second reports the same for the test phase. For all of the datasets, we see a right skew in the histogram of correct responses in the testing phase, indicating that most learners are providing correct answers after training. The skewness for each dataset is -0.45 for Butterflies, -0.81 for Eyes, and -0.37 for Greebles. The skewness is correlated to the amount of improvement learners showed on a given dataset. The numbers imply that the Greebles task is the most difficult and the learners demonstrate the least improvement. The Eyes dataset, on the other hand, is clearly the easiest. We provide additional analysis in the supplementary material (Sec. C).

4.2 Implementation Details

All models we are considering primarily consist of a feature extractor and a classification function. The feature extractor is implemented as a CNN with eight layers (two convolutional, two max-pool, four linear), and is the same across all models (Sec. B). The feature extractor produces a 16 dimensional embedding for an input image which is then processed by the respective classification function. We evaluated one of our models on higher dimensional feature spaces, but found no impact on performance (Sec. A.7). For the two recurrent models, ϕ_{direct} and ϕ_{cls_pred} , we use a three layer LSTM-based [16] fully connected network with a hidden dimension of size 128. The output of the LSTM is passed to a small two-layer network that transforms the output into the desired representation (either a response or a classifier (see Sec. 3.3)). We provide a detailed description of the architectures and inputs in the supplementary material (Sec. F.2).

All models are trained using mini-batch stochastic gradient descent with a batch size of 16 using the Adam optimizer [22]. For the feature extractor, we use a learning rate of 1e-5, and we use a learning rate of 1e-3 for each of the different classification functions. We train using a cross entropy loss. Models are trained with early stopping. Training ends when the best validation loss does not improve for 35 epochs. The upper limit on the number of training epochs is 400. Data is split into train (70%), validation (13.3%), and test (16.7%) splits at the learner level, i.e. sequences from the same individual learner cannot be in more than one split. To ensure more robust results, we validate models by re-shuffling the data five times and report the averages across the splits using micro and macro average precision.

4.3 Results

Tracing Human Learners. We compute the precision and recall curves for each of the visual knowledge tracing models on the held-out learners and report

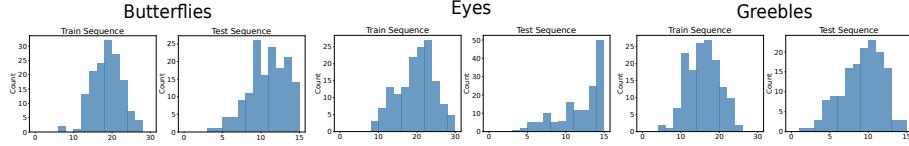


Fig. 3: Human learner performance on our three datasets. For each dataset we provide histograms of learner performance on the respective training and testing sequences. The training results are always worse as it include responses from all time-steps, including when the learner has just started the task and is unfamiliar with the classes. For ‘Greebles’, the worse performance on the test set compared to the other datasets indicates that learners find this task more challenging.

the average precision for each. We also include one additional model for completeness, the ground truth baseline model (GT Label). This baseline does not fit any parameters and simply predicts the corresponding ground truth label of the image for all learners at all time-steps (instead of predicting the learner’s response). Results are summarized in Table 1. Standard deviations are between 0.01 and 0.03 for all models, and do not change the interpretation of the results.

We observe that the recurrent models, ϕ_{direct} and ϕ_{cls_pred} , out-perform the baselines in tracing the learner on the Butterflies ($\sim 8\%$) and Eyes datasets ($\sim 6\%$). However, on the Greebles dataset, the static tracing model, ϕ_{static} , outperforms the recurrent models by ($\sim 3\%$). Both recurrent models are comparable in terms of performance, indicating the reduction in computation described in Sec. 3.3 does not come with a reduction in performance. The time-sensitive tracing model, ϕ_{static_time} , is clearly the worst at tracing learners, likely owing to the unrealistic assumptions it makes about them.

Finally we explore the differences between how different models trace human learners. Fig. 4 shows the average probability of predicting an image correctly for each time-step conditioned on each class for the Butterflies dataset. In the top row, we present the training and test-split average accuracy on each of the five classes over time for the human learners. As previously noted, the training sequences contains 30 randomly selected images and the test sequence contains 15 images. We can see that on average, some classes are much easier than others. In the bottom row, we average model-predicted probabilities for ~ 50 images in each class. To produce the probabilities for the recurrent ϕ_{cls_pred} model, we processes the sequential data of the same set of learners in the top panel. The static tracing model, ϕ_{static} , estimates a hyperplane for each class that roughly tracks the average probability of being correct per class. As expected, this model is not capable of capturing any learning behavior. In contrast, ϕ_{cls_pred} more faithfully traces how the average probability evolves over time. Note, that we process each of the test images independently for ϕ_{cls_pred} .

Table 1: Performance of different visual knowledge tracing approaches on data from human learners. We observe that our two recurrent based models, the direct response ϕ_{direct} and the classifier prediction ϕ_{cls_pred} , perform best on the Butterflies and Eyes dataset but are worse on the synthetic Greebles task. Learners found the Greebles task the most challenging, and as a result, there was much less learning occurring compared to the first two datasets. ‘GT Label’ is an additional baseline that uses the corresponding ground truth class label y as the prediction of the learner’s response r .

	Greebles				Eyes				Butterflies			
	Train		Test		Train		Test		Train		Test	
	Micro	Macro										
GT Label	0.48	0.51	0.58	0.61	0.56	0.56	0.69	0.69	0.45	0.44	0.50	0.49
ϕ_{static}	0.63	0.52	0.67	0.58	0.60	0.59	0.67	0.68	0.55	0.53	0.64	0.61
ϕ_{static_time}	0.52	0.44	0.49	0.40	0.34	0.34	0.33	0.34	0.54	0.52	0.61	0.59
ϕ_{direct}	0.70	0.59	0.77	0.64	0.66	0.65	0.75	0.74	0.55	0.53	0.60	0.57
ϕ_{cls_pred}	0.71	0.62	0.77	0.65	0.65	0.65	0.74	0.74	0.54	0.52	0.60	0.57

4.4 Discussion

Comparing Models. We observe that our recurrent models are quite effective at tracing human learner knowledge on visual categorization tasks for datasets where there is a clear learning signal. On the Greebles dataset, which is the most challenging and displays the least amount of learning, we see that the simple static tracing model ϕ_{static} is less prone to over-fitting and is thus marginally better. The time-sensitive tracing model ϕ_{static_time} performs the worst overall. Unlike the recurrent methods, it is unable to share information between time-steps, forcing it to fit a classifier at each time-step with only ~ 90 training points (the number of learners in the training set after we split out the validation and test sets). This makes it extremely prone to over-fitting on the limited training data that is available.

The direct response model ϕ_{direct} and the classifier prediction model ϕ_{cls_pred} differ in their outputs and their inputs. ϕ_{direct} includes \mathbf{z} , the representation of the query image \mathbf{x} to be classified by the learner, as input to the LSTM. ϕ_{cls_pred} outputs weights of a multi-class linear classifier that is used to classify the representation \mathbf{z} . Unlike ϕ_{direct} , \mathbf{z} is not an input to the LSTM for ϕ_{cls_pred} . However, both models incorporate the ground truth label y for \mathbf{x} as input. These structural differences between the models make little difference to the tracing performance, but the ϕ_{cls_pred} model is more efficient if it needs perform evaluation multiple times for new query items. In supplementary experiments (A.1), we explore the impact of removing the ground truth class label y for the query image from the input. We observe that this results in a large decrease in performance for both models, suggesting that this class information is valuable to the model when making future predictions. Also in the supplement, we compare two cognitive models (prototype and exemplar [31]) (A.3), a Transformer model [11] (A.4), a pre-trained ResNet feature extractor (A.5), and input meta-information to the tracing models (A.6). We find that the cognitive, Transformer, and ResNet

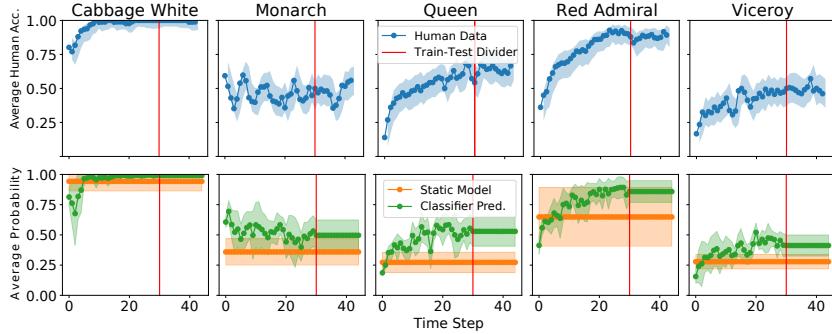


Fig. 4: (**Top**) The smoothed average human learner accuracy for class over time from the Butterflies dataset. The shadowed regions indicate confidence intervals as the number of samples in each time and class bin are not guaranteed to be the same. (**Bottom**) The average probability of having a class correctly predicted by the static ϕ_{static} model (orange) and the recurrent ϕ_{cls_pred} model (green). At each time-step, for each learner in the test set, the models predict class probabilities for ~ 50 images per class. The probabilities are averaged (solid line) and the shadows indicate one standard deviation. In both rows, the red line indicates the point at which the learners switch from training to testing. After that, the models will continue to produce the same probabilities on the test images for the remaining time-steps as the sampled images do not change.

models do not out-perform the recurrent architectures, but are worth further exploration. Additionally, including meta-information in the input vectors results in a performance increase across all models we tested, suggesting this a promising direction for future work. Finally, we explore the representations learned by the recurrent models (Secs. D, E).

Limitations and Future Work. Currently, we train our feature encoder f from scratch for each task on relatively small amounts of image data. One source of improvement would be to pre-train the feature space so that it better reflects human visual similarity judgements. Such an improved feature space would provide a better starting point for task specific finetuning. Replacing the LSTM with an appropriately designed Transformer network [47] is another change that could result in greater flexibility for the model. Transformers are better able to capture long-term dependencies and have been shown to be useful in knowledge tracing on non-visual educational datasets [33,11]. However, it remains to be seen if this would be valuable for visual knowledge tracing.

Our approaches do not explicitly model memory decay – the phenomenon of memory ‘fading’ due to the passage of time [41,17]. This is likely to be more of a problem when tracing over longer time horizons, e.g. days or weeks, as opposed to the multiple minute long sessions that our learners engage in. Similarly, given the short time durations of our teaching sessions, we assume that no significant

‘feature learning’ is happening for individual learners. Instead we model learners as attending to a subset of the different visual features that are captured in our joint embedding space. In future work, it would be interesting to further explore if these two assumptions are valid.

Applications. Successful tracing of human learners has implications for crowd-sourcing annotations, metric learning, and machine teaching. Early detection of poor annotators in crowd sourcing would reduce monetary and time costs in labelling datasets. Additionally, identifying annotators with ‘specialist’ knowledge could allow for targeted crowd sourcing, tailoring to the abilities of each individual annotator. A successful tracing algorithm should be able to predict future performance with increasing confidence as the learners are being trained on the annotation task.

The most impactful domain for a successful tracing algorithm is automated teaching, e.g. teaching medical image interpretation skills [10,2]. Teaching humans is challenging because their knowledge state is unobserved, it changes over time, and the information they provide using current interfaces can be limited. In this work, we show that with a reasonable amount of training data (i.e. data from ~ 150 learners), and only a single response at each time-step, we are able to capture information about the learner’s current knowledge for visual classification tasks. The types of visual knowledge tracing approaches presented could be used in conjunction with machine teaching methods. The more a teaching algorithm knows about the learner, the more effective it can be when selecting examples to present to them.

5 Conclusion

More accurate models of human visual classification will lead to improved methods for crowd annotation collection, better techniques for automatically teaching visual knowledge to human learners, and perhaps provide us with insight into how we can build future artificial systems that are more data efficient. To this end, in this work we explored the problem of visual knowledge tracing – the task of predicting the internal, potentially time varying, image classification function used by human learners. To do this, we presented a series of models that range in complexity from basic static linear classifiers all the way to recurrent models that take a learner’s prior response history into account when making predictions about their future behavior. We collected new annotations for three challenging visual classification tasks from humans engaged in a visual learning task in order to benchmark the performance of these different models. Our results show that our recurrent neural network-based models resulted in the most faithful reproductions of unobserved learner predictions on real image datasets. Finally, we outlined limitations of our work and pointed to open questions that require further investigation.

Acknowledgements: Thanks to the anonymous reviews for their valuable feedback. This work was in part supported by the Turing 2.0 ‘Enabling Advanced Autonomy’ project funded by the EPSRC and the Alan Turing Institute and also by the Simons Collaboration on the Global Brain.

References

1. Prolific, www.prolific.co, accessed Mar 7 2022
2. Amiri, E., Sha, P., Palmer, E.M.: Training novices to discriminate retinal diseases using perceptual learning. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. pp. 1456–1460 (2020)
3. Ashby, F.G., Maddox, W.T.: Human category learning. *Annu. Rev. Psychol.* **56**, 149–178 (2005)
4. Ashby, F.G., Maddox, W.T.: Human category learning 2.0. *Annals of the New York Academy of Sciences* (2011)
5. Ashby, F.G., O'Brien, J.B.: Category learning and multiple memory systems. *Trends in cognitive sciences* **9**(2), 83–89 (2005)
6. Attarian, M., Roads, B.D., Mozer, M.C.: Transforming neural network visual representations to predict human judgments of similarity. *arXiv:2010.06512* (2020)
7. Barry, D.N., Love, B.C.: Human learning follows the dynamics of gradient descent. *PsyArXiv* (2021)
8. Biederman, I., Shiffrar, M.M.: Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, memory, and cognition* **13**(4), 640 (1987)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
10. Cheng, C.T., Chen, C.C., Fu, C.Y., Chaou, C.H., Wu, Y.T., Hsu, C.P., Chang, C.C., Chung, I., Hsieh, C.H., Hsieh, M.J., et al.: Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging* **11**(1), 1–8 (2020)
11. Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., Heo, J.: Towards an appropriate query, key, and value computation for knowledge tracing. In: Proceedings of the Seventh ACM Conference on Learning@ Scale (2020)
12. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? In: CVPR (2022)
13. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* **4**(4), 253–278 (1994)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
15. Hebart, M.N., Zheng, C.Y., Pereira, F., Baker, C.I.: Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature human behaviour* **4**(11), 1173–1185 (2020)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
17. Hunziker, A., Chen, Y., Mac Aodha, O., Gomez Rodriguez, M., Krause, A., Perona, P., Yue, Y., Singla, A.: Teaching multiple concepts to a forgetful learner. *NeurIPS* (2019)
18. Johns, E., Mac Aodha, O., Brostow, G.J.: Becoming the expert-interactive multi-class machine teaching. In: CVPR. pp. 2616–2624 (2015)
19. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. *Symmetry* (2019)
20. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5), 1122–1131 (2018)

21. Kim, K.H., Mac Aodha, O., Perona, P.: Context embedding networks. In: CVPR. pp. 8679–8687 (2018)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
23. Kruskal, J.B., Wish, M.: Multidimensional scaling. No. 11, Sage (1978)
24. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. IJCV (2020)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
26. Liu, Q., Shen, S., Huang, Z., Chen, E., Zheng, Y.: A survey of knowledge tracing. arXiv:2105.15106 (2021)
27. Mac Aodha, O., Su, S., Chen, Y., Perona, P., Yue, Y.: Teaching categories to human learners with visual explanations. In: CVPR. pp. 3820–3828 (2018)
28. Markman, E.M.: Categorization and naming in children: Problems of induction. mit Press (1989)
29. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: ECCV. pp. 681–699 (2020)
30. Nigam, I., Tokmakov, P., Ramanan, D.: Towards latent attribute discovery from triplet similarities. In: ICCV. pp. 402–410 (2019)
31. Nosofsky, R.M., Meagher, B., Kumar, P.: Contrasting exemplar and prototype models in a natural-science category domain. In: CogSci (2020)
32. Nosofsky, R.M., Sanders, C.A., Meagher, B.J., Douglas, B.J.: Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods **50**(2), 530–556 (2018)
33. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv:1907.06837 (2019)
34. Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., Sohl-Dickstein, J.: Deep knowledge tracing. NeurIPS (2015)
35. Pu, S., Yudelson, M., Ou, L., Huang, Y.: Deep knowledge tracing with transformers. In: International Conference on Artificial Intelligence in Education. pp. 252–256 (2020)
36. Roads, B.D., Love, B.C.: Enriching imagenet with human similarity judgments and psychological embeddings. In: CVPR (2021)
37. Roads, B.D., Mozer, M.C.: Predicting the ease of human category learning using radial basis function networks. Neural Computation (2021)
38. Roth, K., Millich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning. In: ICML. pp. 8242–8252 (2020)
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
40. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015)
41. Settles, B., Meeder, B.: A trainable spaced repetition model for language learning. In: ACL (2016)
42. Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., Krause, A.: Near-optimally teaching the crowd to classify. In: ICML. pp. 154–162 (2014)
43. Tamuz, O., Liu, C., Belongie, S.J., Shamir, O., Kalai, A.: Adaptively learning the crowd kernel. In: ICML (2011)

44. Tan, R., Vasileva, M.I., Saenko, K., Plummer, B.A.: Learning similarity conditions without explicit supervision. In: ICCV. pp. 10373–10382 (2019)
45. Van Der Maaten, L., Weinberger, K.: Stochastic triplet embedding. In: International Workshop on Machine Learning for Signal Processing. pp. 1–6 (2012)
46. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: CVPR (2015)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
48. Veit, A., Belongie, S., Karaletsos, T.: Conditional similarity networks. In: CVPR. pp. 830–838 (2017)
49. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR. pp. 1386–1393 (2014)
50. Wang, P., Nagrecha, K., Vasconcelos, N.: Gradient-based algorithms for machine teaching. In: CVPR. pp. 1387–1396 (2021)
51. Wang, P., Vasconcelos, N.: A machine teaching framework for scalable recognition. In: ICCV. pp. 4945–4954 (2021)
52. Welinder, P., Branson, S., Perona, P., Belongie, S.: The multidimensional wisdom of crowds. NeurIPS (2010)
53. Wilber, M., Kwak, I., Belongie, S.: Cost-effective hits for relative similarity comparisons. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 2, pp. 227–233 (2014)
54. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
55. Xu, F., Tenenbaum, J.B.: Word learning as bayesian inference. Psychological review (2007)
56. Zeithamova, D.: Category learning systems. The University of Texas at Austin (2008)
57. Zhu, X., Singla, A., Zilles, S., Rafferty, A.N.: An overview of machine teaching. arXiv:1801.05927 (2018)

A Additional Experiments

In this section, we present several additional models and also consider the impact of embedding dimension on performance.

A.1 RNN Variants

We can vary the input information to both of the recurrent models in three ways. The notation in parentheses maps to the entries in the later supplementary tables.

(base): The models only receive the history of images, ground truth class labels, and learner responses

$$\psi_{rnn}(\mathbf{z}_{1:t}^k, y_{1:t}^k, r_{1:t}^k)). \quad (7)$$

(y): In addition to the history, the model receives the ground truth class label of the image shown to the learner at the current time-step

$$\psi_{rnn}(\mathbf{z}_{1:t}^k, y_{1:t}^k, r_{1:t}^k, y)). \quad (8)$$

(y, z): Finally, as in the main paper, the model can include both the ground truth class of the image and the representation of the image from the learned CNN

$$\psi_{rnn}(\mathbf{z}_{1:t}^k, y_{1:t}^k, r_{1:t}^k, \mathbf{z}, y)). \quad (9)$$

The results of the variants are presented in Tables 2, 3, and 4.

A.2 DKT

Next, we adapt Deep Knowledge Tracing (DKT) [34] to our setting. We deviate from the original DKT method in two main ways. First, the types of queries (e.g. math problems) in educational datasets do not allow for instance-level representations. Instead, skills (i.e. question types) were jointly encoded with information about whether the problem was answered correctly by the learner. Second, the output of DKT was the learner’s probability of being correct for each skill, not a particular question instance.

We modify the DKT algorithm to make it appropriate for the setting described in our work. We replace skills with the class-level label for an image and convert the output into a probability distribution over the class labels such that it can be trained with the cross-entropy loss

$$\phi_{dkt}(y) = \sigma(\psi_{rnn}(y_{1:t}^k, r_{1:t}^k, y))). \quad (10)$$

At a high-level, this model variant encodes no instance-level (i.e. image) information to make its predictions.

We observe that this DKT model (ϕ_{dkt}) performs slightly worse in all cases, indicating that image information is valuable to enable the models to better trace learner performance. The results of the DKT variant are presented in Tables 2, 3, and 4.

A.3 Cognitive Models

Cognitive models make stronger assumptions on how humans learn. We modify the prototype and exemplar models described in the cognitive science literature [31] and evaluate them on our datasets.

Prototype. The prototype model proposes that learners store a prototypical image for each class. Each new image is compared to the learner’s class prototypes and the highest similarity class is selected. In our formulation, the class prototype is the average feature representation of previously seen images of that class. In the following equations, τ is the current time-step, $P_\tau^k(c)$ is the prototype of class c for learner k at time-step τ , and δ is the dirac-delta function and acts as a selector for images from class c ,

$$P_\tau^k(c) = \frac{1}{\tau - 1} \cdot \sum_t^{t-1} z_t^k * \delta(y_\tau^k - c), \quad (11)$$

$$\hat{r}_\tau^k(c) = \frac{\text{sim}(P_\tau^k(c), z_\tau^k)}{\sum_c \text{sim}(P_\tau^k(c), z_\tau^k)}. \quad (12)$$

Exemplar. The exemplar model proposes that learners store previously seen images in a memory bank of exemplars. Query images are compared to all of the exemplars. The learner chooses the class with the highest total similarity to the query image. In the following equations, $E_\tau^k(c)$ is the sum of the class c similarity scores for learner k at time-step τ with respect to the current image z_τ^k . Following [31], we introduce a learnable parameter γ to scale the similarities (this value is fixed to 1 in the prototype model),

$$E_\tau^k(c) = \sum_t^{t-1} \text{sim}(z_t^k, z_\tau^k) \cdot \delta(y_\tau^k - c), \quad (13)$$

$$\hat{r}_\tau^k(c) = \frac{\text{sim}(E_\tau^k(c), z_\tau^k)^\gamma}{\sum_c \text{sim}(E_\tau^k(c), z_\tau^k)^\gamma}. \quad (14)$$

Both models compute similarity by using an exponential decay function over the Euclidean distance between feature representations of the images,

$$\text{sim}(z_i, z_j) = e^{-c*d(z_i, z_j)}. \quad (15)$$

Finally, instead of learning the feature space separately with visual similarity experiments, we jointly estimate a CNN along with the model parameters to discover the feature space.

We find that these models perform worse than the models presented in the main paper. However, simple modifications (like weighting the history of exemplars or images in the prototype) may help. Exploring the space of cognitive models is an interesting direction for future work. The results of these variants are presented in Tables 2, 3, and 4.

A.4 Transformers

Recently, the knowledge tracing community has found the Transformer architecture to be an effective model for tracing human learners in non-visual tasks. We modify the SAINT model [11] for our visual learning setting. First, we introduce a CNN-based feature extraction stage to embed images. The encoder receives the current image’s embedding and its ground truth label. The decoder gets the previous learner response. The decoder predicts the learner’s response to the image (also passed to the encoder).

The Transformer model does surprisingly poorly on these datasets. We expect that future work exploring Transformer architectures designed for this task will demonstrate

performance on par with the recurrent models. The results of the Transformer model are presented in Tables 2, 3, and 4.

A.5 ResNet Backbone Experiments

We swap out our CNN backbone with a ResNet-18 [?] pre-trained on ImageNet. We freeze the weights in layers 1, 2 and 3, but leave layer 4 to be learned. The output of layer 4 is passed to a fully-connected layer that reduces the output of the layer to the desired dimensionality, as opposed to the final classifier used for the ImageNet classification task. The results of these experiments are presented in Table 5.

A.6 Including Per-Class Accuracy as Input

We find that including some meta-information can help with tracing performance. To do this, we compute a learner’s accuracy on each class at each time-step and concatenate this vector to the input of three tracing models (ϕ_{static} , ϕ_{direct} and ϕ_{cls_pred}). We find a boost in performance across all models. The results are presented in Table 6, where we observe a boost in performance. It is likely that other sources of meta-information (such as time-taken on an example) will also help [?].

A.7 Varying Embedding Dimension

We demonstrate that varying the embedding dimension of the feature extractor has little effect on the performance of the direct response model (Table 7).

Table 2: Performance of all model variants on the Butterflies dataset. The model variant is denoted in the subscript corresponding to the same subscripts in A.1. One can see that $\phi_{direct(base)}$ performs poorly for a recurrent model. This model does not have access to any information about the current time-step and is effectively guessing both the image that will be shown and the associated response. We also show the per-class average precision scores on the train sequence in addition to the micro and macro scores from before. These scores show that the benefit of the recurrent models appear primarily in classes that have large changes in average performance (e.g. Red Admiral) over the training period. The models with \dagger are models presented in Table 1 of the main paper. The scores are reported with their standard deviations and the top average performers in each column are in bold.

	Butterflies								
	Train				Test				
	Cabbage White	Monarch	Queen	Red Admiral	Viceroy	Micro	Macro	Micro	Macro
GT Label \dagger	0.94 \pm 0.03	0.32 \pm 0.01	0.36 \pm 0.04	0.65 \pm 0.04	0.27 \pm 0.01	0.48 \pm 0.02	0.51 \pm 0.02	0.58 \pm 0.02	0.61 \pm 0.02
$\phi_{static}\dagger$	0.95 \pm 0.01	0.37 \pm 0.02	0.34 \pm 0.06	0.66 \pm 0.03	0.27 \pm 0.03	0.63 \pm 0.02	0.52 \pm 0.02	0.67 \pm 0.02	0.58 \pm 0.03
$\phi_{static_time}\dagger$	0.97 \pm 0.01	0.34 \pm 0.03	0.29 \pm 0.03	0.35 \pm 0.03	0.24 \pm 0.02	0.52 \pm 0.02	0.44 \pm 0.01	0.49 \pm 0.01	0.40 \pm 0.01
ϕ_{dkt}	0.95 \pm 0.02	0.43 \pm 0.02	0.45 \pm 0.04	0.72 \pm 0.05	0.33 \pm 0.06	0.67 \pm 0.02	0.57 \pm 0.02	0.74 \pm 0.02	0.64 \pm 0.02
$\phi_{transformer}$	0.96 \pm 0.01	0.36 \pm 0.03	0.34 \pm 0.02	0.69 \pm 0.06	0.26 \pm 0.01	0.62 \pm 0.02	0.52 \pm 0.02	0.68 \pm 0.05	0.58 \pm 0.04
$\phi_{prototype}$	0.95 \pm 0.02	0.32 \pm 0.03	0.30 \pm 0.05	0.59 \pm 0.08	0.27 \pm 0.03	0.53 \pm 0.03	0.48 \pm 0.03	0.63 \pm 0.01	0.54 \pm 0.02
$\phi_{exemplar}$	0.90 \pm 0.03	0.30 \pm 0.03	0.27 \pm 0.04	0.35 \pm 0.13	0.26 \pm 0.03	0.44 \pm 0.05	0.42 \pm 0.04	0.53 \pm 0.07	0.45 \pm 0.07
$\phi_{direct(base)}$	0.34 \pm 0.02	0.29 \pm 0.03	0.23 \pm 0.02	0.26 \pm 0.02	0.22 \pm 0.03	0.27 \pm 0.01	0.27 \pm 0.01	0.20 \pm 0.01	0.20 \pm 0.01
$\phi_{direct(y)}$	0.97 \pm 0.02	0.43 \pm 0.03	0.48\pm0.05	0.76 \pm 0.07	0.38\pm0.04	0.71\pm0.02	0.60 \pm 0.02	0.78\pm0.02	0.66\pm0.01
$\phi_{direct(y, z)}\dagger$	0.97 \pm 0.01	0.41 \pm 0.04	0.44 \pm 0.06	0.77 \pm 0.07	0.36 \pm 0.03	0.70 \pm 0.03	0.59 \pm 0.02	0.77 \pm 0.03	0.64 \pm 0.03
$\phi_{cls_pred(base)}$	0.96 \pm 0.01	0.41 \pm 0.06	0.32 \pm 0.05	0.59 \pm 0.14	0.25 \pm 0.02	0.59 \pm 0.06	0.51 \pm 0.05	0.61 \pm 0.06	0.52 \pm 0.05
$\phi_{cls_pred(y)}\dagger$	0.98\pm0.01	0.46\pm0.02	0.48\pm0.04	0.78\pm0.05	0.38\pm0.02	0.71\pm0.02	0.62\pm0.01	0.77 \pm 0.02	0.65 \pm 0.02
$\phi_{cls_pred(y, z)}$	0.98\pm0.01	0.45 \pm 0.01	0.47 \pm 0.04	0.78\pm0.05	0.37 \pm 0.02	0.70 \pm 0.02	0.61 \pm 0.01	0.77 \pm 0.03	0.66\pm0.02

Table 3: Performance of all model variants on the Eyes dataset. Please see the caption of Table 2 for more details.

	Eyes							
	Train		Test					
	DME	Drusen	Normal	Micro	Macro	Micro	Macro	
GT Label†	0.56±0.02	0.54±0.03	0.58±0.02	0.56±0.02	0.56±0.02	0.69±0.02	0.69±0.01	
$\phi_{static}†$	0.63±0.03	0.53±0.05	0.62±0.02	0.60±0.03	0.59±0.03	0.67±0.02	0.68±0.02	
$\phi_{static_time}†$	0.32±0.01	0.35±0.01	0.36±0.01	0.34±0.00	0.34±0.01	0.33±0.01	0.34±0.02	
ϕ_{dkt}	0.63±0.02	0.60±0.03	0.65±0.02	0.63±0.02	0.63±0.02	0.74±0.03	0.73±0.03	
$\phi_{transformer}$	0.41±0.09	0.41±0.05	0.42±0.09	0.41±0.08	0.41±0.08	0.37±0.02	0.39±0.02	
$\phi_{prototype}$	0.61±0.04	0.50±0.04	0.56±0.04	0.56±0.03	0.56±0.03	0.65±0.04	0.65±0.05	
$\phi_{exemplar}$	0.57±0.02	0.48±0.02	0.59±0.03	0.54±0.02	0.55±0.02	0.68±0.04	0.67±0.04	
$\phi_{direct(base)}$	0.38±0.02	0.40±0.01	0.38±0.02	0.38±0.01	0.39±0.01	0.35±0.01	0.34±0.01	
$\phi_{direct(y)}$	0.65±0.03	0.62±0.03	0.68±0.03	0.66±0.02	0.65±0.02	0.75±0.01	0.73±0.02	
$\phi_{direct(y,z)}†$	0.64±0.02	0.62±0.03	0.69±0.02	0.66±0.02	0.65±0.02	0.75±0.01	0.74±0.02	
$\phi_{cls_pred(base)}$	0.48±0.05	0.39±0.02	0.48±0.05	0.45±0.04	0.45±0.03	0.44±0.02	0.44±0.02	
$\phi_{cls_pred(y)}†$	0.65±0.02	0.62±0.03	0.69±0.01	0.65±0.03	0.65±0.02	0.74±0.02	0.74±0.04	
$\phi_{cls_pred(y,z)}$	0.64±0.01	0.62±0.03	0.69±0.02	0.65±0.02	0.65±0.02	0.75±0.01	0.74±0.02	

Table 4: Performance of all model variants on the Greebles dataset. Please see the caption of Table 2 for more details.

	Greebles							
	Train		Test					
	Agara	Bari	Cooka	Micro	Macro	Micro	Macro	
GT Label†	0.51±0.02	0.37±0.03	0.43±0.03	0.45±0.02	0.44±0.02	0.50±0.01	0.49±0.01	
$\phi_{static}†$	0.63±0.03	0.43±0.04	0.55±0.05	0.55±0.03	0.53±0.03	0.64±0.01	0.61±0.01	
$\phi_{static_time}†$	0.64±0.04	0.39±0.02	0.54±0.04	0.54±0.03	0.52±0.03	0.61±0.01	0.59±0.02	
ϕ_{dkt}	0.59±0.03	0.41±0.03	0.49±0.05	0.52±0.02	0.50±0.02	0.59±0.02	0.55±0.02	
$\phi_{transformer}$	0.52±0.11	0.36±0.03	0.45±0.05	0.46±0.07	0.45±0.06	0.44±0.08	0.43±0.08	
$\phi_{prototype}$	0.58±0.03	0.42±0.01	0.54±0.05	0.52±0.02	0.51±0.02	0.58±0.02	0.57±0.02	
$\phi_{exemplar}$	0.59±0.02	0.43±0.03	0.52±0.05	0.52±0.03	0.51±0.03	0.63±0.01	0.60±0.01	
$\phi_{direct(base)}$	0.37±0.02	0.35±0.02	0.36±0.01	0.36±0.00	0.36±0.00	0.34±0.02	0.34±0.02	
$\phi_{direct(y)}$	0.59±0.02	0.41±0.03	0.51±0.04	0.52±0.02	0.50±0.03	0.59±0.02	0.55±0.02	
$\phi_{direct(y,z)}†$	0.62±0.03	0.42±0.03	0.55±0.05	0.55±0.03	0.53±0.03	0.60±0.02	0.57±0.03	
$\phi_{cls_pred(base)}$	0.63±0.03	0.40±0.02	0.56±0.06	0.55±0.02	0.53±0.03	0.61±0.02	0.60±0.03	
$\phi_{cls_pred(y)}†$	0.62±0.03	0.41±0.03	0.54±0.03	0.54±0.02	0.52±0.03	0.60±0.02	0.57±0.03	
$\phi_{cls_pred(y,z)}$	0.63±0.03	0.41±0.03	0.55±0.03	0.55±0.02	0.53±0.01	0.61±0.02	0.59±0.02	

Table 5: Performance of models trained using a pre-trained ResNet with partially frozen weights (as described in Sec. A.5). We only compare model variants that appear in the main text. Similar to the original experiment results the classifier prediction model (ϕ_{cls_pred}) performs the best. However, the overall performance decreases slightly across the board. We observe a larger decrease for the direct response model (ϕ_{direct}), likely due to the larger dependence that it has on the feature space.

	Butterflies								
	Train				Test				
	Cabbage White	Monarch	Queen	Red Admiral	Viceroy	Micro	Macro	Micro	Macro
GT Label	0.94±0.03	0.32±0.01	0.36±0.04	0.65±0.04	0.27±0.01	0.48±0.02	0.51±0.02	0.58±0.02	0.61±0.02
ϕ_{static}	0.95±0.01	0.36±0.02	0.32±0.04	0.65±0.04	0.27±0.03	0.62±0.02	0.51±0.02	0.67±0.03	0.57±0.03
ϕ_{static_time}	0.44±0.02	0.25±0.02	0.22±0.01	0.33±0.03	0.19±0.02	0.28±0.01	0.29±0.00	0.27±0.02	0.34±0.02
ϕ_{direct}	0.97±0.01	0.41±0.03	0.38±0.06	0.74±0.07	0.28±0.03	0.66±0.03	0.55±0.02	0.70±0.03	0.58±0.04
ϕ_{cls_pred}	0.97±0.01	0.39±0.02	0.49±0.02	0.75±0.06	0.32±0.03	0.69±0.01	0.59±0.01	0.75±0.01	0.65±0.02

Table 6: Performance of models after concatenating per-class accuracy information to the input vector for the tracing model (Sec. A.6). We only compare model variants that appear in the main text. We observed a boost for all models.

	Greebles				Eyes				Butterflies			
	Train		Test		Train		Test		Train		Test	
	Micro	Macro										
ϕ_{static}^\dagger	0.63	0.52	0.67	0.58	0.60	0.59	0.67	0.68	0.55	0.53	0.64	0.61
$\phi_{static+perClAcc}$	0.65	0.54	0.69	0.58	0.63	0.62	0.69	0.70	0.59	0.57	0.66	0.65
ϕ_{direct}^\dagger	0.70	0.59	0.77	0.64	0.66	0.65	0.75	0.74	0.55	0.53	0.60	0.57
$\phi_{direct+perClAcc}$	0.71	0.62	0.78	0.67	0.69	0.69	0.78	0.78	0.53	0.51	0.61	0.57
$\phi_{cls_pred}^\dagger$	0.71	0.62	0.77	0.65	0.65	0.65	0.74	0.74	0.54	0.52	0.60	0.57
$\phi_{cls_pred+perClAcc}$	0.71	0.61	0.79	0.67	0.69	0.69	0.79	0.79	0.53	0.51	0.60	0.58

Table 7: We train the ϕ_{direct} tracing model on the butterflies dataset with different embedding dimensions. We find that embedding dimension has no impact on performance.

	Butterflies								
	Train				Test				
	Cabbage White	Monarch	Queen	Red Admiral	Viceroy	Micro	Macro	Micro	Macro
ϕ_{direct_dim8}	0.98±0.01	0.41±0.03	0.45±0.06	0.77±0.07	0.35±0.03	0.70±0.03	0.59±0.03	0.77±0.03	0.65±0.04
ϕ_{direct_dim16}	0.98±0.01	0.42±0.05	0.46±0.04	0.77±0.07	0.37±0.03	0.70±0.02	0.60±0.02	0.78±0.03	0.66±0.03
ϕ_{direct_dim32}	0.98±0.01	0.43±0.04	0.46±0.03	0.77±0.06	0.35±0.02	0.70±0.02	0.60±0.01	0.78±0.02	0.67±0.02
ϕ_{direct_dim64}	0.97±0.01	0.43±0.03	0.48±0.04	0.78±0.06	0.37±0.03	0.71±0.02	0.61±0.02	0.79±0.01	0.66±0.01

B CNN Architecture Details

In Table B8 we describe the architecture of the CNN used to encode images for all of the models.

Table B8: Structure of the CNN backbone used to learn the image representation. The bolded and italicized entries are variable and depend on the experiment and dataset. The number of image channels (*img_chns*) is three for the Butterflies and Greebles dataset, but is one for Eyes. The Butterflies and OCT datasets contain larger images (144 x 144), and so *img_feats* is set to 1296. For the Greebles dataset, the images are (128 x 128) and *img_feats* is set to 1204. Finally, the output of the model is the size of the embedding dimension and is set to 16 for all experiments.

CNN backbone						
layer	in channels	out channels	k	s	p	activation
conv1	<i>img_chns</i>	8	5	1	2	PReLU
maxpool1	-	-	4	-	-	PReLU
conv2	8	16	5	1	2	PReLU
maxpool2	-	-	4	-	-	-
flatten	-	-	-	-	-	-
linear	<i>img_feats</i>	512	-	-	-	PReLU
linear	512	256	-	-	-	PReLU
linear	256	256	-	-	-	PReLU
linear	256	16	-	-	-	PReLU

C Additional Results

We recreate Fig. 4 for all datasets and include results from the Direct Response and Time-Sensitive Model. For the Greebles dataset, we include the histograms of the features of the classes to demonstrate the difficulty of the task.

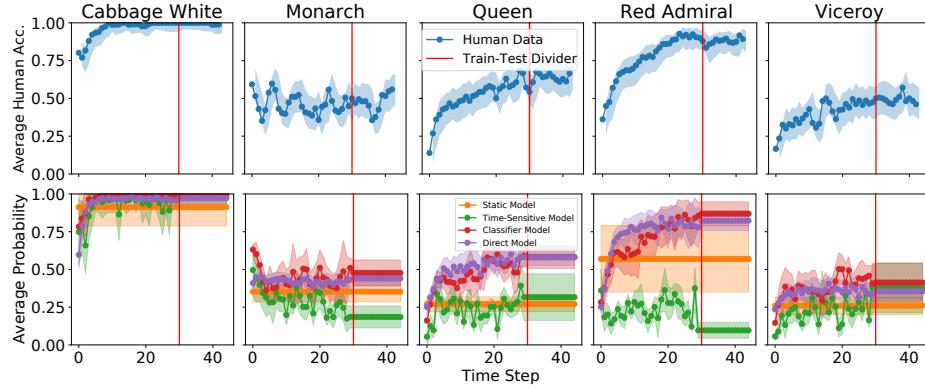


Fig. C5: **(Top)** The smoothed average human learner accuracy for each class over time on the Butterflies dataset. The shadowed regions indicate confidence intervals as the number of samples in each time and class bin are not guaranteed to be the same. **(Bottom)** The average probability of getting a class correct predicted by the static ϕ_{static} model (orange), ϕ_{static_time} model (green), the direct response ϕ_{direct} model (purple), and the classifier prediction ϕ_{cls_pred} model (red). At each time-step, for each learner in the test set, the models predict class probabilities for ~ 50 images per class. The probabilities are averaged (solid line) and the shadows indicate one standard deviation. While both recurrent models have similar traces, the ϕ_{direct} produces smoother average probabilities.

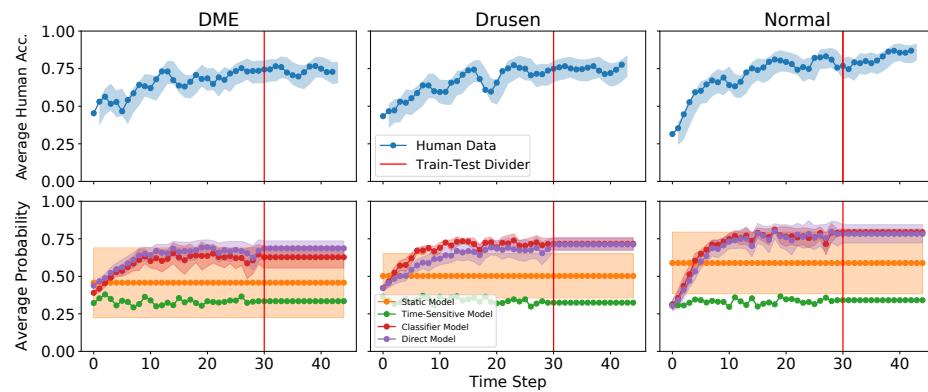


Fig. C6: Human and model performance on the Eyes dataset. See Fig. C5 for a detailed caption.

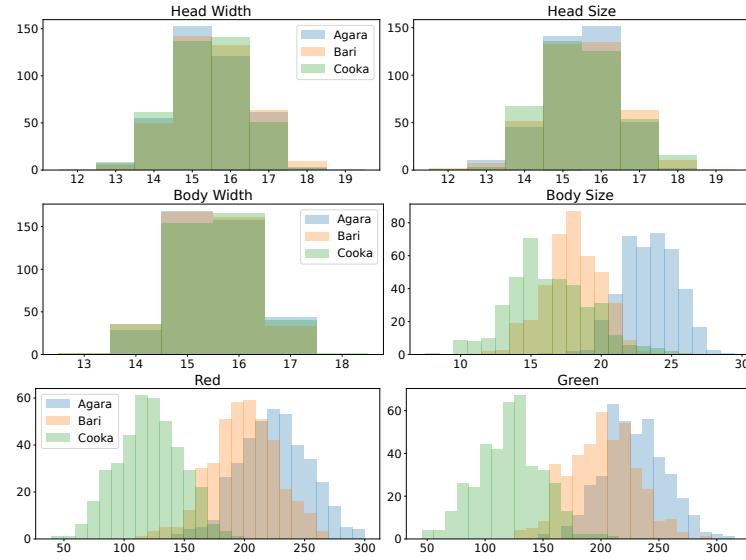


Fig. C7: The Greebles dataset was inspired by the one used in [52]. In our version, the three classes vary in Head Width and Size (top row), Body Width and Size (middle row), and the Red and Green channel for the RGB color (bottom row). The histograms overlap completely for Head Width, Head Size, and Body Width. These variations serve as distractors since they provide no information about which class an image belongs to. The other features, Body Size, the Red channel, and the Green channel have different distributions and can be used to estimate the class. Agara and Bari are most separable by Body Size, Cooka is most separable from both Agara and Bari in the two color channels. However, note that they are not perfectly separated and it is possible, although less likely, for two images from different classes to take on the same properties. This makes the Greebles dataset particularly challenging, since the important features are both subtle and imperfect for distinguishing between classes.

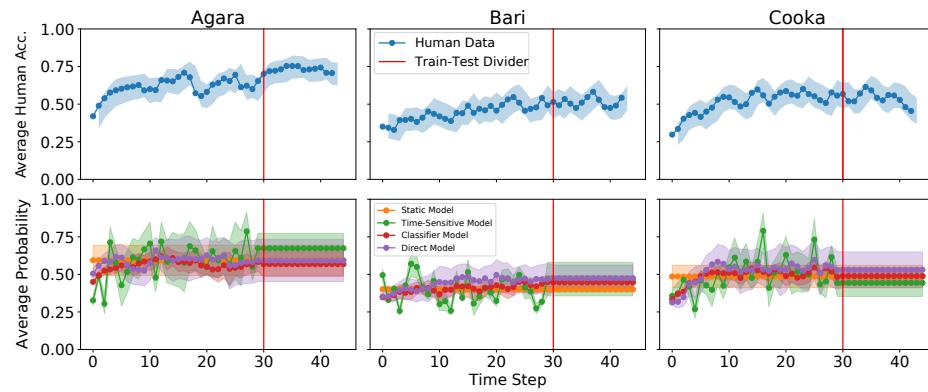


Fig. C8: Human and model performance on the Greebles dataset. See Fig. C5 for a detailed caption.

D Learned Representations

Here we explore the representations learned by the classifier prediction model (ϕ_{cls_pred}) on the Butterflies dataset. In Fig. D9 we visualize the internal state of the model and in Fig. D10 we provide an in depth comparison for two different learners.

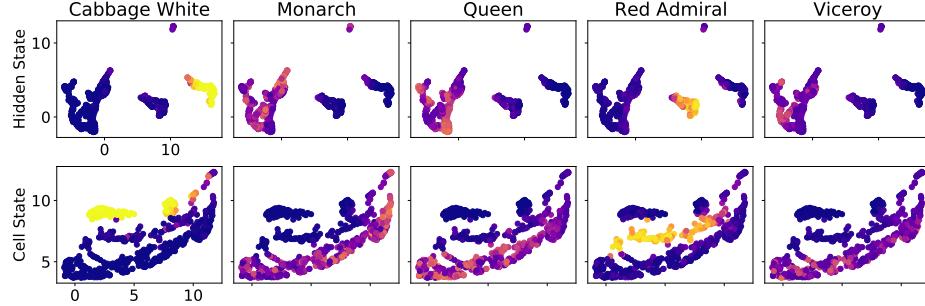


Fig. D9: The hidden states and cell states of the LSTM for ϕ_{cls_pred} while tracing 25 test-set learners are plotted in 2D using the UMAP dimensionality reduction algorithm [?]. **(Top)** The hidden state representations are colored according to the probability (purple to yellow) that a response produced with that hidden state would correctly classify an image of the class in the panel title. We can see that the classes that correspond to the best average performance by humans are in well-defined clusters (e.g. Cabbage White), whereas the classes that are commonly mistaken for each other are grouped together and have much weaker probabilities of being correct. **(Bottom)** The cell states are visualized in the same manner. For the cell states, we can see that the clusters seem to be dragged across a single dimension. The Cabbage White and Red Admiral cluster is split in two pieces in the cell state, which we explore in Fig. D10.

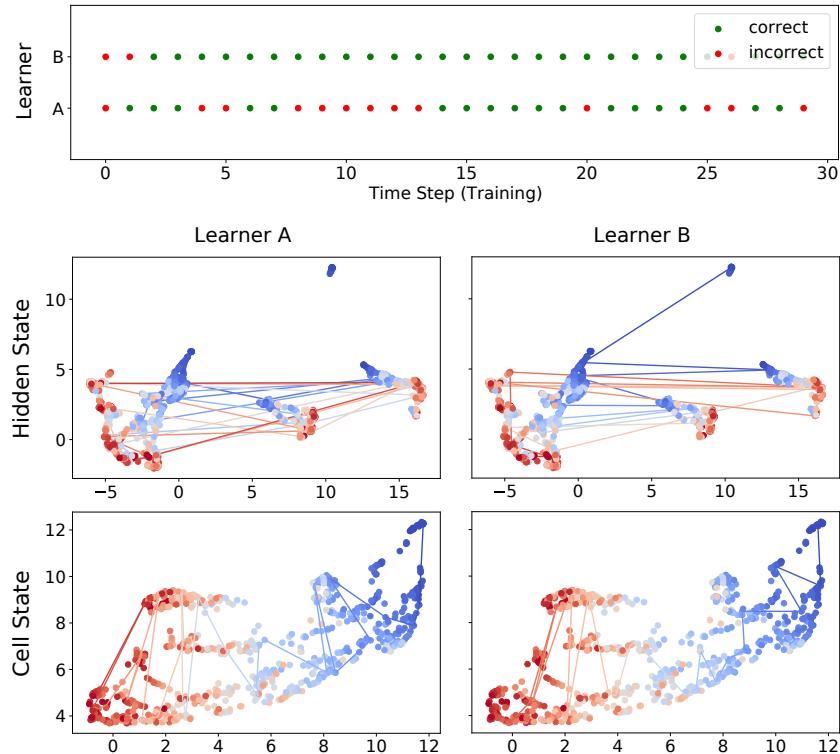


Fig. D10: **(Top)** The sequence of correct and incorrect responses made by two human learners during training. We selected these two learners as they demonstrate different learning behaviors. It seems Learner B may already be familiar with butterflies. **(Bottom)** We overlay each learners’ trajectory through the hidden and cell states. The colors represent the time-step, where dark blue is the beginning of training, light-grey is the middle, and dark red is the end. We see that Learner B’s trajectory quickly skips to the left of the cell state, suggesting the LSTM encodes the learners skill level on all classes in certain dimensions of the cell state and uses the hidden state to translate the skill level into an appropriate response for the image shown to the learner.

E Feature Space

In Fig. E11 we visualize the feature space learned by the CNN for the classifier prediction model (ϕ_{cls_pred}) on the Butterflies dataset.

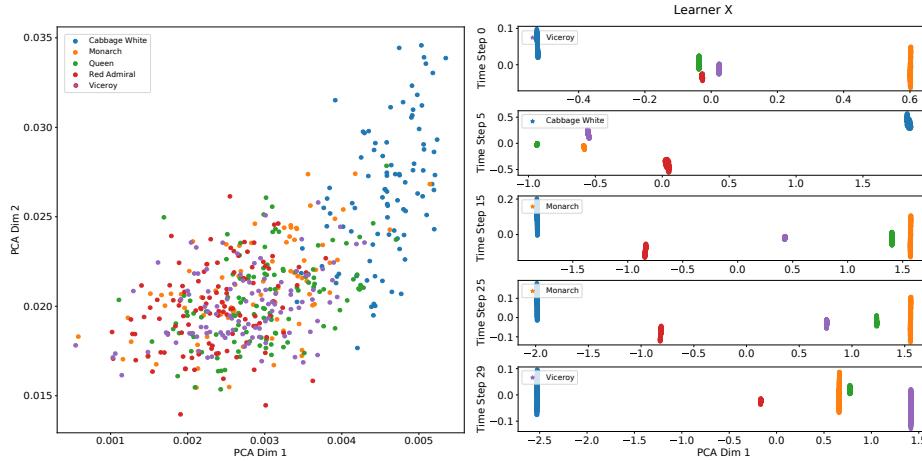


Fig. E11: The feature space learned by the CNN must support several types of behaviors, since behavior changes between learners and over time. We use PCA to reduce the learned feature space into two dimensions. **(Left)** We show a subset of images in the Butterflies dataset colored by the ground truth label. Aside from the Cabbage White class, which is the easiest to identify, the representations are difficult to separate. **(Right)** We use the hyperplanes predicted while tracing a single learner X to induce a subspace and visualize the features in that subspace. Within the subspaces, the classes are much better separated. Each row shows a subspace induced by a hyperplane for different time-steps - where the time-step is indicated on the left. The colors represent the class and the labelled color is the target class for the image being evaluated in that time-step. We see that, over time, the target class is pushed further to the right and is better separated from the other classes (see orange cluster in time-step 15 vs. 25). Classes that are confused for each other have less separation, whereas classes like Cabbage White, that are rarely confused, are extremely well-separated from the other classes. Also, note that the subspace orientation (target class moved to the right) matches how the dot product between the hyperplanes and features is translated into probabilities in the model.

F Additional Implementation Details

F.1 Types of Learner Responses

There are several ways to request information from the learner: they can provide their best guess, a ranked list of guesses, or confidence scores for each class. In these datasets, we ask for a ranking of each learner’s top 3 classes as a balance between time-spent and informativeness. While our models are trained on their top choice (equivalent to their best guess), we hypothesize that the extra information available in the ranked responses can be leveraged to improve response prediction performance. We leave this to future work.

F.2 Recurrent Neural Networks

Here we elaborate on the details of the recurrent neural network based models.

Direct Response Model. At each time-step, this model receives the hidden states, cell states, the learner’s response to the previous interaction, the embedding of the current image, and the true label of the current image. The model predicts the response of the learner with respect to the input image.

Classifier Prediction Model. At each time-step, this model receives the hidden state, cell state, the embedding of the image from the previous interaction, the learner’s response to the previous interaction, and the true label of the current image. The model predicts a classifier that is used to classify the embedding of the input image such that it matches the response of the learner to that input image at that time-step.

Ground truth labels and learner responses are represented as one-hot-encoded vectors. For both models, at the first time-step, some of the values that make up the input vector are not available to the model. For example, the hidden state, cell state, the response to previous interaction, etc. These vectors are initialized to zeros.