Neural concept learning: Self-attention models for zero-likelihood observations

Jerry Mao (jerrym@mit.edu)
Massachusetts Institute of Technology, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Abstract

Concept learning is a problem that has been well-studied through the lens of Bayesian frameworks. However, these frameworks fail to model human cognition when presented with observations from novel concepts. We propose a new model for concept learning based on self-attention networks, removing explicit Bayesian constraints to provide a model even when the observation has zero likelihood. We provide theoretical foundations showing that neural models can exhibit Bayesian-like behavior and demonstrate it in practice. We also evaluate our model on human datasets, finding that these models do not yet exhibit out-of-sample generalization and explainability.

Keywords: concept learning; Bayesian models; self-attention

Introduction

Human cognition is a field that has been widely studied through the lens of Bayesian models. One problem used to test theories of cognition is concept learning. Bayesian frameworks for concept learning have been found to perform very well at modelling human behavior (Tenenbaum, 1999).

However, these models can lack the ability to handle instances where the prior probability or observation likelihood are zero. While humans can react online to observations from novel concepts, these examples cause the posterior distribution to be undefined, rendering Bayesian inference impossible. In this work, we study an alternative method for concept learning that does not impose a Bayesian framework.

In particular, we explore a novel model of concept learning based on self-attention neural networks (Vaswani et al., 2017). We draw from the intuition that self-attention can learn pairwise correspondence between elements, and suggest encoding concepts in latent representations. We find that such a model behaves similarly to a Bayesian model, although its predictions given zero-likelihood observations are often uninterpretable. Our specific contributions are:

- A new method for concept learning based on self-attention neural networks.
- An evaluation of the model's performance relative to established Bayesian models.
- A study of the model's interpretations of human behavior, especially in cases where Bayesian models fail.

All source code for this work is available open-source at `https://github.com/j-mao/neural-concept-learning`.

Problem statement

We first formalize the problem of concept learning as it is used in our study.

Within some universe $\mathcal{U}$, we define a concept to an abstract rule, specifying a set $C \subseteq \mathcal{U}$ of elements that satisfy it. Some examples of concepts could include "even numbers" or "numbers between 11 and 20" in a universe of positive integers. The concept learning task is to, given a subset of randomly sampled elements $D \subseteq C$ (the "observation"), estimate the membership of elements in the concept C itself.

Tenenbaum (1999) studies this problem using a Bayesian framework. Specifically, they propose to estimate the posterior probability according to Bayes' theorem:

$$\Pr[C \mid D] \propto \Pr[D \mid C] \Pr[C] \tag{1}$$

$$\Pr[x \in C \mid D] = \textstyle\sum_{C_i : x \in C_i} \Pr[C_i \mid D] \tag{2}$$

Notably, the constant of proportionality in Eq. (1) is $\Pr[D]^{-1}$, which can be undefined if D is not consistent with any concept in the concept space. In these situations, the likelihood of D is zero under all known concepts. While humans may still produce some estimate for the posterior, this Bayesian model is unable to model human cognition in any capacity.

In our work, we propose an alternative model for the concept learning problem, based on self-attention networks (Vaswani et al., 2017). This model does not impose a Bayesian structure and therefore does not encounter the same issue. Therefore, we aim to answer the following questions:

- How accurate is a neural model of concept learning at modelling human behavior? How is model capacity and accuracy affected by the latent dimension?
- Can a neural model of concept learning exhibit emergent Bayesian behavior?
- What interpretations can a neural model assign to human responses on the concept learning problem, including on observations from novel concepts outside the known concept space?

Method

In this section we reintroduce the *Number game* as a task for concept learning, and describe our neural architecture for our modelling framework.

Table 1: The space of concepts used in our work. Each concept had a predetermined size for the observation set generated from it.

Concept rule	Size $\lvert C\rvert$	Observation $\lvert D\rvert$
Multiples of 2	15	4
Multiples of 3	10	4
Multiples of 4	7	3
Multiples of 5	6	3
Between 1 and 10	10	4
Between 11 and 20	10	4
Between 21 and 30	10	4
Nonsense	N/A	3

Task: Number game

Tenenbaum (1999) uses the *Number game* as a task for evaluating models of concept learning. We also use this task for our work.

In the *Number game*, the universe is the set of positive integers between 1 and N, and concepts are a restricted set of mathematical properties of these integers. In the original work, $N = 100$; however, here we set $N = 30$ to reduce the size of a representative human dataset. Table 1 lists the specific concepts used in our experiment.

Human data collection

A human dataset was collected by asking volunteers to solve the number game. A collection of 27 observation sets D_i was generated and divided evenly into 3 "rooms" as shown in Table 2. Each participant was presented with a room and asked to provide a set of numbers $\tilde{C}_i$ for each observation, representing their best guess for the underlying concept.

Of the 9 observations in each room, 2 were "nonsense" observations composed entirely of random numbers, thus falling outside the concept space. Experiment participants were oblivious to the existence of these nonsense concepts. The question order was randomly shuffled to ensure these observations were interspersed.

Participants were invited to solve multiple rooms if they wished. The division of rooms is superfluous and only present as a load-balancing mechanism; each room received a similar distribution of concepts.

In order to establish the participants' priors, they were asked to read a cover story before being presented with the observations. The cover story is included in the Appendix.

Neural network architecture

We model the task with a simple self-attention network f equipped with an n-dimensional latent space. We represent each integer x by a "query" vector $q_x \in \mathbb{R}^n$ and key vector $k_x \in \mathbb{R}^n$. For each set of numbers $D \subseteq \mathcal{U}$ from a concept C, the model predicts *relative* probabilities as follows.

$$f(x \mid D;\, q,k) \propto \exp\left(\sum_{z\in D} \langle q_x, k_z\rangle\right) \tag{3}$$

Table 2: The observations presented to participants, annotated with the room it appeared in and the ground truth concept it was drawn from.

Room	Observation D	Ground-truth concept
A	12, 18, 26, 28	Multiples of 2
A	4, 10, 20, 28	Multiples of 2
B	2, 14, 16, 30	Multiples of 2
C	6, 8, 22, 24	Multiples of 2
A	6, 9, 21, 27	Multiples of 3
B	3, 9, 15, 21	Multiples of 3
C	12, 15, 18, 24	Multiples of 3
A	4, 12, 16	Multiples of 4
B	8, 24, 28	Multiples of 4
C	4, 8, 20	Multiples of 4
B	10, 20, 25	Multiples of 5
C	5, 15, 30	Multiples of 5
A	4, 5, 6, 9	Between 1 and 10
A	1, 2, 3, 7	Between 1 and 10
C	1, 5, 6, 8	Between 1 and 10
A	11, 14, 16, 19	Between 11 and 20
B	12, 15, 19, 20	Between 11 and 20
B	11, 13, 17, 18	Between 11 and 20
B	22, 25, 27, 29	Between 21 and 30
C	21, 23, 28, 30	Between 21 and 30
C	22, 24, 26, 29	Between 21 and 30
A	6, 9, 19	Nonsense
A	3, 14, 23	Nonsense
B	1, 5, 13	Nonsense
B	6, 11, 27	Nonsense
C	3, 7, 15	Nonsense
C	11, 24, 25	Nonsense

Eq. (3) is implemented using softmax attention. We do not estimate the probability $\Pr[x \in C \mid D]$ itself, but we estimate that it is proportional to the output $f(x \mid D)$. Note that $f(\cdot \mid D)$ is not a probability distribution over $\mathcal{U}$.

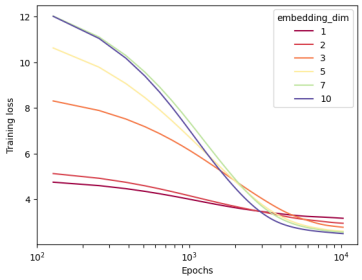
We learn these embeddings via gradient-based optimization using a negative log-likelihood objective on the sets $\tilde{C}_i$, omitting data from nonsense observations.

$$J(q,k) = \mathbb{E}_i\left[\sum_{x\in\tilde{C}_i} -\ln f(x \mid D_i;\, q,k)\right] \tag{4}$$

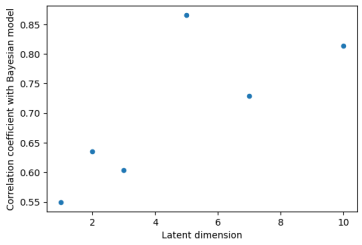
We emphasize that Eq. (4) trains only on positive examples $x \in \tilde{C}_i$. Regardless, training with this objective function naturally decreases the estimated probabilities for $\mathcal{U} \setminus \tilde{C}_i$ relative to these positive examples.

We additionally train a reference network using a collection of 1024 observations drawn directly from the concepts. The objective function for this network uses the ground truth $\tilde{C}_i = C_i$.

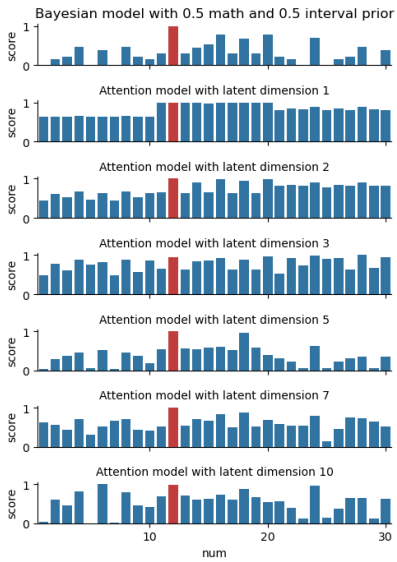
We compare these two networks with established Bayesian frameworks and test their emergent properties on our datasets.



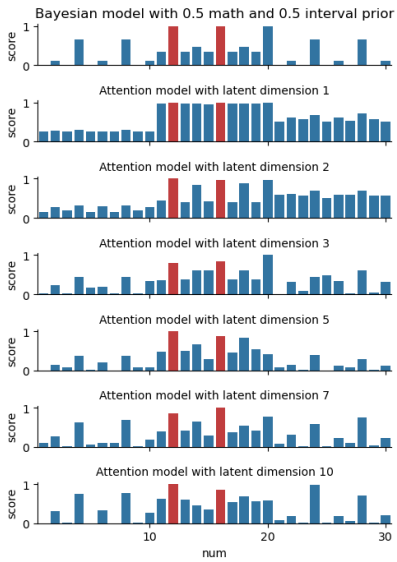
(a) Learning curve for attention models.



(b) Correlation with Bayesian model for $D = \{12\}$



(c) Predictions for $D = \{12\}$



(d) Predictions for $D = \{12, 16\}$

Figure 1: Comparison of attention models with varying latent dimensions, alongside a baseline Bayesian model with uniform priors over the categories. Elements of D are shown in red. Predictions are relative to the maximum $f(\cdot \mid D)$ in each trial.

Results

In this section we present the results of our experiments. We train each network with Adam (Kingma & Ba, 2015) over 10240 passes through their respective datasets.

We additionally compare our networks with a baseline Bayesian model equipped with the following likelihood function suggested by Tenenbaum (1999).

$$\Pr[D \mid C] = |C|^{-|D|} \tag{5}$$

We use Eq. (5) in the framework defined in Eqs. (1) and (2), with a uniform prior over the two categories of concepts ("multiples" and "intervals"), distributed uniformly within each category.

Model capacity and emergent behavior

We first study the properties of our neural model when trained on the ground-truth objective. We train models with $n = 1$, 2, 3, 5, 7, or 10 latent dimensions in order to understand model capacity.

Fig. 1a shows that with enough training time, the networks with a larger latent dimension achieve a lower loss J. After approximately 3000 epochs, the widest network with $n = 10$ achieves a loss lower than all the other networks.

More interestingly, Figs. 1c and 1d examine the actual effect of reducing the latent dimensionality on the model's inferences. With just one latent dimension, there is enough representation power for only one feature; we see that the model has learned to categorize the universe into three sets, corresponding to the three interval concepts.

The model appears to learn a second feature when a second latent dimension is added. When $n = 2$, the inferred probabilities appear to gain an "alternating" pattern between low and high values, representing each number's parity. This corresponds with the "multiple of 2" concept.

We note that the nature of Eq. (3) means that the neural model never infers a zero probability, even when the concept space makes certain elements impossible. However, we argue that it is this exact property which allows it to robustly handle novel situations where the observation is impossible under the concept space. Nevertheless, the model does learn to minimize these values as it predicts near-zero values.

Indeed, we qualitatively see that once there are at least 5 dimensions, the neural model's inferences begin to correlate very strongly with the Bayesian model. Fig. 1b illustrates a generally increasing trend in the correlation coefficient as n increases, reaching over 0.80. Moreover, comparing between Figs. 1c and 1d we see the inferred probabilities change plausibly as a new element is added to the observation D.

In fact, we argue that the structure of Eq. (5) enables this behavior by design. We elaborate on this further in the Discussion section.

Human dataset

In this section we pivot to studying the neural model trained on human-collected data. A total of 132 responses were collected, distributed as 51 from Room A, 40 from Room B and 41 from Room C. Collectively they yielded 924 predictions $\tilde{C}_i$ for "sensible" observations (ie., those with ground-truth concepts), and a further 264 predictions for nonsense observations.

Fig. 2 shows an example of the model's inferences when it is trained on the sensible observations. The model's perfor-

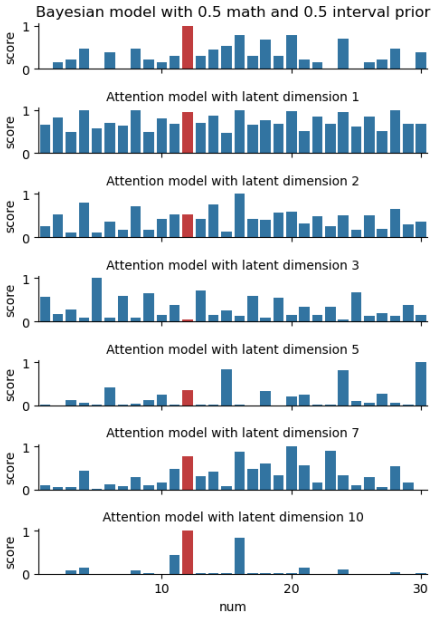
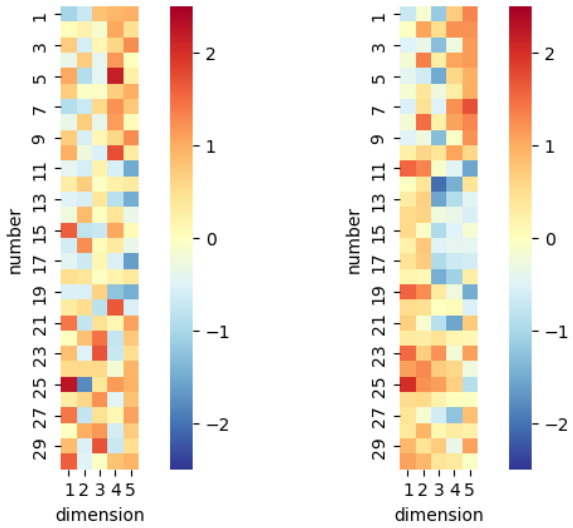


Figure 2: Predictions for $D = \{12\}$ when trained on human data.



(a) Query embeddings q (b) Key embeddings k

Figure 3: Learned embedding vectors using the ground-truth training objective when $n = 5$.

mance is significantly degraded compared to the one trained from ground truth data: understandably so, as human data is generally noisy. It appears that the model with latent dimension 7 is close to mimicking the Bayesian model; however, this behavior is not robust.

In fact, sometimes the model even predicts extremely low probabilities for elements that belonged to the observation D. This was seen when the latent dimension was 3.

Due to the poor quality of this model, we perform no further analysis on it.

Model interpretations

In this section we inspect our model in order to interpret the reasons both for model inferences and for human inferences. We do so by examining any relationships between human predictions and the learned parameters of the model trained with the ground-truth objective.

We begin by studying those parameters in isolation. Fig. 3 visualizes these parameters for the model with $n = 5$, which had achieved the highest correlation with the Bayesian model.

There are some concepts that appear to be very clearly present in the embedding. For example, dimension 5 consists of generally positive values for numbers in $[1,10]$ and $[21,30]$, and negative values for numbers in $[11,20]$. This very clearly encodes the interval concepts.

Beyond this, it is difficult to immediately find an obvious interpretation in terms of the concepts at first glance. Further inspection reveals that each dimension may in fact be a blend of several concepts. For instance, dimension 4 has very high query values for numbers in the "multiples of 5" concept, while also appearing to use negative numbers for the interval concepts. Several key embeddings appear to be encoding parity for numbers up to 10, but diverge to encoding other information for larger numbers.

As such, viewing the parameters gives us some limited amount of insight into model explainability.

Finally, we test our model's predictive power on a nonsense observation shown in Fig. 4. The observation $D = \{1,5,13\}$ is consistent with a hypothetical concept "odd numbers", but is nonsense to the model as this concept is not part of the original concept space. As previously discussed, the Bayesian framework fails to produce any prediction for this nonsense observation.

On the other hand, our neural model does produce an output for this case; unfortunately, it does not appear to be very interpretable. Despite "odd numbers" being the complement of the seen concept "multiples of 2", the predictions from the model did not exhibit any obvious meaning. It appears that the neural model has poor out-of-sample generalization to these out-of-sample nonsense observations.

Indeed, Table 3 shows model underperformance on the

Table 3: Evaluation loss on the human dataset for model trained on ground-truth objective.

	Human dataset type	
Latent dimension	Sensible	Nonsense
$n = 1$	3.564	3.643
$n = 2$	3.687	4.270
$n = 3$	3.671	4.767
$n = 5$	3.550	4.688
$n = 7$	3.662	5.088
$n = 10$	3.771	5.725

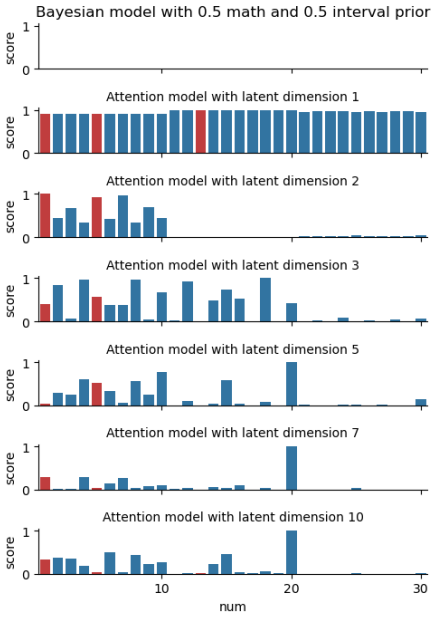


Figure 4: Model inferences for a nonsense observation with $D = \{1, 5, 13\}$.

nonsense dataset. We also see symptoms of overfitting as loss increases with latent dimension. These may all contribute to challenges in achieving generalization.

Discussion

Our empirical results suggest that the neural model is able to mimic some of the behaviors of a Bayesian agent. In this section, we argue that this is possible due to the architecture of the model itself.

We begin focusing our attention on the following toy scenario: consider a space of two concepts $C_1 \subset C_2 \subseteq \mathcal{U}$ with equal priors, with an observation $D \subseteq C_1$. Let $x \in C_1$ and $y \in C_2 \setminus C_1$. We can compute under the Bayesian framework:

$$\begin{aligned} \frac{\Pr[x \in C \mid D]}{\Pr[y \in C \mid D]} &= \frac{\Pr[D \mid C_1] + \Pr[D \mid C_2]}{\Pr[D \mid C_2]} \\ &= \frac{|C_1|^{-|D|} + |C_2|^{-|D|}}{|C_2|^{-|D|}} \\ &= 1 + \left(\frac{|C_2|}{|C_1|}\right)^{|D|} \end{aligned} \tag{6}$$

On the other hand, suppose the neural model f has $n = 2$ and learns embeddings $q_z, k_z \in \mathbb{R}^2$ such that for some constant $c > 1$, $(q_z)_i = (k_z)_i = [\![z \in C_i]\!]\sqrt{\ln c}$, where $[\![\cdot]\!]$ denotes the Iverson bracket. Then,

$$\begin{aligned} \frac{f(x \mid D)}{f(y \mid D)} &= \exp\left(\sum_{z \in D} \langle q_x, k_z \rangle - \langle q_y, k_z \rangle\right) \\ &= \exp\left(|D| \ln c\right) \\ &= c^{|D|} \end{aligned} \tag{7}$$

Examining Eqs. (6) and (7) shows that for the appropriate value of c, the neural model behaves just like the Bayesian model as $|D| \to \infty$.

As such, the design of the network architecture enables some Bayesian-like behavior.

We remark that this exact analysis may not generalize to more complex scenarios. The lack of an enforced Bayesian structure also limits the model's capabilities. For example, when $x \notin C_2$ it will not infer a zero probability like the Bayesian model, and the above analysis does not produce the same result. Moreover, even when $x \in D \subseteq C$ is part of the observation, the model does not always predict high values for $\Pr[x \in C \mid D]$.

Despite these limitations, we empirically see that our model nevertheless learns to behave in a seemingly Bayesian manner. A fruitful line of further work may be in improving model interpretability, in order to better understand why the model is successful in certain cases, as well as its failure modes. We suggest exploring constraints that may discourage embedding dimensions from blending concepts, as well as methods for analyzing embeddings that may extract the concepts as "principle components".

We note that additional insight may be gained from further inspection of the network's learned parameters. The exponent in Eq. (3) can be rephrased as the inner product with a sum of key vectors:

$$\sum_{z \in D} \langle q_x, k_z \rangle = \left\langle q_x, \sum_{z \in D} k_z \right\rangle \tag{8}$$

As such, it may be fruitful to examine summed key vectors for observation sets, alongside query vectors for elements in the human response sets.

Conclusion

We propose a model for concept learning based on self-attention networks as a means of handling exceptional cases where observations do not fit in any known concept. We demonstrate that these neural models can produce predictions for probabilities of concept membership even when Bayesian models fail under zero-likelihood scenarios. We further show that the neural architecture admits Bayesian behavior in limiting cases, and that this behavior can even be observed on regular datasets.

We remark that these neural models suffer from generalization challenges and may exhibit poor interpretability on these out-of-sample zero-likelihood inputs. However, our analysis shows that model parameters may alleviate interpretability issues: we highlight that it may be possible to recover concepts from these latent representations, and recommend further study of these parameters to bring additional insight to interpretability.

Acknowledgements

Computation power for this work was funded by Google Cloud education credits.

References

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations.*

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems.*

Appendix: Cover story

Included below is the cover story used to establish the priors of experiment participants. The story was intended to encourage participants to think mathematically, but also to prefer simpler rules due to the mechanical nature of the machines.

> You are an archeologist, and you have just discovered 3 rooms of relics from an ancient civilization. The civilization was profoundly knowledgeable, and knew about many properties of mathematics. They were also expert mechanical engineers, and have created many great machines.
>
> One type of machine is a number "classifier"; these machines take a number as input, and determine whether that number follows a secret rule. Each machine could have a different secret rule, and the machines are independent.
>
> Each of the 3 rooms contains 9 of these machines; each machine has a single dial you can turn, to select an input integer between 1 and 30. After playing with the machines, you've already found some numbers that each machine "accepts" as part of its secret rule. Can you guess some other numbers that also follow each machine's secret rule?