Predictions emerge in neural networks trained to perceive Bach's music

Akanksha Gupta (emailakankshagupta@gmail.com)
Aix-Marseille University, Inserm, Institut de Neurosciences des Systèmes (INS), Marseille, France

Alejandro Tabas (tabas@bcbl.eu)
Perceptual Inference Group, Basque Center on Cognition, Brain and Language, San Sebastian, Spain
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Abstract

Predictive processing proposes that the prior knowledge relevant for inference is compressed into a prediction on the immediately future states. Here we inquire whether neural networks trained to infer the current latent state in musical sequences develop a set of internal predictions on what comes next.

We used noisy tokenized Bach compositions as sensory inputs and trained RNN as models of neural circuits. We first trained the networks to infer the current latent state (token of the composition without noise) given a stream of observations (tokens of the composition with noise). After the training, we inspected whether the internal states of the network stored predictive information on the next token. To do this, we fitted a linear readout from the hidden states of the network optimized to predict the next latent state. To ensured that the predictions were stored in the network and not computed by the linear readout, we compared the predictive performance of the network with that of a linear network trained to predict the next latent state based on the current latent state.

The results confirm that neural circuits optimized to perceive the current state learn to predict future sensory input, suggesting that predictive capabilities emerge as a natural consequence of such optimization. These findings offer computational evidence for predictive processing and provide insights into how biological systems might compress their prior knowledge and use it to navigate in noisy environments.

Keywords: Predictive Processing; Bayesian Brain Hypothesis; Recurrent Neural Networks (RNNs); Emergent Behaviors

The brain makes use of its prior knowledge to accurately infer the state of the external world from noisy and ambiguous sensory information. Predictive processing proposes that the prior knowledge relevant for inference is compressed into a prediction on the immediately future states (Friston, 2003; Aitchison & Lengyel, 2017). The sensory input is then combined with the predictions using Bayesian believe updating. Although this strategy is optimal when inferring latent states from simple stochastic systems, predictions might not be useful when inferring latent states from more complex generators like music or language.

Here we inquire whether neural networks trained to infer the current latent state in musical sequences develop the capacity to predict what comes next. If that was the case, this would be a strong indicator that, even when the generators are too complex to perform optimal inference using predictions, neural circuits may still rely on predictions to infer the state of the external world.

Methods

We used noisy tokenized Bach compositions as sensory inputs. A latent token x_t represented the actual note or chord of one of Bach's composition at time t as a multi-label one-hot encoding (Figure 1A, red barplot). Observation tokens y_t were sampled from a Gaussian distribution with mean x_t and a scalar-matrix variance (Figure 1A, yellow barplot).

We used recurrent neural networks (RNNs) as models of neural circuits. RNNs were first trained to infer the latent token x_t based on the series of noisy observations $y_{1:t}$ (Figure 1A).

Networks were trained to minimize the mean binary cross-entropy (BCE) across all tokens in each sequence. We trained 168 RNNs with a varying number of hidden units (2 to 256) and observation noises (standard deviation between 10^{-3} and 2.0).

To measure whether the network was exploiting contextual information to infer the latent states, we compare the network performance with that of a baseline model (a one-step feed-forward network with the same number of hidden units as the RNN) trained on the same loss and receiving as inputs only the current observation y_t.

After the training, we inspected whether the internal states of the network stored predictive information on the next token. To do this, we fitted a linear readout (Figure 1A, blue lines) from the hidden states of the network optimized to predict x_{t+1} after the network had processed $y_{1:t}$.

To ensure that the predictions were stored in the network and not computed by the linear readout, we compared the predictive performance of the network with that of a linear network trained to predict x_{t+1} based on x_t (low-bound Markovian model). Our reasoning was that, if predictions are driven solely from the internal states of the network, the accuracy should be lower or equal to that of a linear network that takes as inputs the actual x_t.

Input tokens were derived from a corpus with 74 Bach compositions in MIDI format. Each composition was converted into a sequence of tokens, where a new token started every time a single note would change in the composition. Loudness and duration information was disregarded so that each token represented either a single note or a chord. Tokens were further compressed into a chromatic representation of

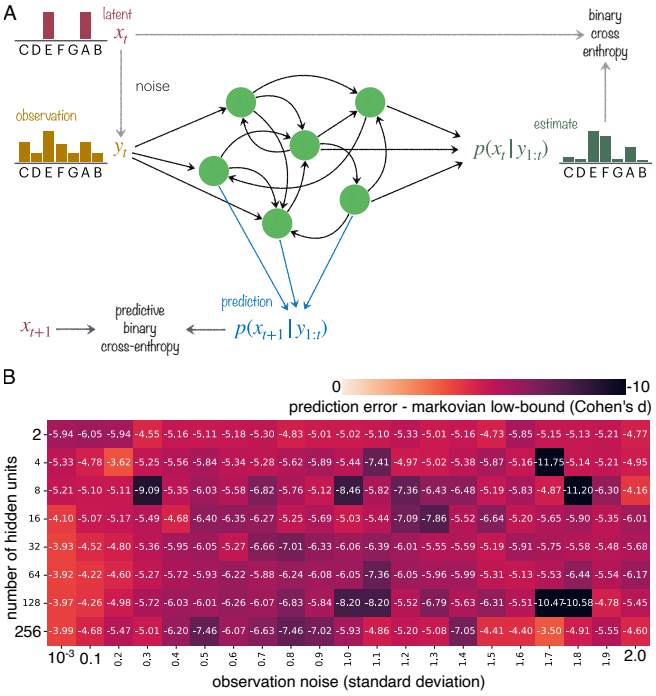

Figure 1: **A) Training architecture and strategy.** The sequences of latent variables $x_{1:t}$ are multi-label one-hot encoding derived from Bach's compositions. Observations y_t are noisy observations of the latent variables x_t. The networks were first trained to minimize binary cross-entropy (BCE) between the target x_t and the output of the network, which is assumed to encode $p(x_t|y_{1:t})$. After training, a linear readout from the network states is trained to minimize BCE between the next target x_{t+1} and the output of the readout, which is assumed to encode $p(x_{t+1}|y_{1:t})$. **B) Results.** We compared the predictive performance of the network against a low-error-bound Markovian model. The heatmap plots the effect size of the difference of BCE between the network and the baseline model as Cohen's d. Negative values indicate that the RNN had a lower error rate (i.e., better performance) than the baseline. Effect sizes are plotted for different observation noise levels (x-axis) and network sizes (y-axis).

12 dimensions. The dataset was partitioned into training (52 compositions), validation (7), and test (15) subsets.

Each RNN consisted of a single-layer gated recurrent unit (GRU) network (Cho et al., 2014). Training was performed using the Adam optimizer with a learning rate of $\eta = 0.02$. Training subset input $y_{1:t}$ and target $x_{1:t}$ sequences where splitted into 512 tokens chunks and arranged in 512 sequence batches. To avoid overfitting, we measured the performance of the network on the validation set every 100 batches; training was early-stopped if the error (i.e., BCE) in the validation set was lower or equal than a cutoff value $\theta = \mu_{traing} - \sigma_{training}$, where μ_{traing} and $\sigma_{training}$ are the mean and standard deviation of the BCE in the training set. After training, performance was measured using the complete testing set.

Results

The RNNs learned to denoise the observations better than the baseline model across all network sizes and noise levels, indicating that they successfully learned to integrate information over time to infer the current token more accurately.

The linear readouts of the RNNs consistently outperformed the baseline model (Figure 1B), demonstrating that the RNNs had implicitly learned to predict the next token. The results indicate that predictions may emerge as a natural byproduct of optimization in networks trained solely on perception.

Conclusion

Perceptual systems need to infer the content of the sensory world from noisy, ambiguous, and incomplete sensory signals. Mounting evidence from behavioral and neurophysiological studies suggests that this inference process is powered by Bayesian believe updating, where prior information is compressed into a predictive distribution that continuously updated according the sensory inputs to conform our perception of reality. Previous authors have suggested that this believe updating is the optimal strategy to to denoise the sensory inputs during perceptual inference (Friston, 2003). However, this point has so-far been mathematically derived for systems that have very specific evolution rules; namely, linear Gaussian dynamic Markovian systems and their generalizations.

Here we showed that RNNs optimized to infer latent states that evolved via a generative model of a much greater complexity also make use of predictions. Although real-world signals such as human speech are often more complex, Bach's compositions show a rich hierarchical structure that can be non-trivially used to form expectations about future states.

The results provide computational support for the predictive processing framework and offer valuable insights into how both artificial and biological systems might use predictions to navigate an environment filled with uncertainty.

Acknowledgments

This project was kickstarted during BrainHack Donostia 2024. We would like to thank BrainHack Donostia and the team who worked in this project during those days. AT is founded by the ERC (SynPrePro, Grant number 101115798)

References

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. doi: 10.1016/j.conb.2017.08.010

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, September). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (No. arXiv:1406.1078). arXiv. doi: 10.48550/arXiv.1406.1078

Friston, K. (2003, November). Learning and inference in the brain. *Neural networks*, *16*(9), 1325–52. doi: 10.1016/j.neunet.2003.06.005