Scalable gaussian process inference of neural responses to movies

Simone Azeglio, Thomas Buffet, Matthew Chalk

Predicting the responses of sensory neurons to natural stimuli is a long-standing goal. Recent work has shown how, in certain cases, deep neural networks can perform well at this task [1]. However, their performance can be substantially degraded when trained on insufficient data, e.g. when recording time is limited. Moreover, to constructing these models requires multiple structural choices (e.g. the network architecture, non-linearities, hyperparameters etc.), and it may be unclear *a priori* how these will affect overall performance. On the other hand, gaussian processes (GPs) require few assumptions and perform well with limited data, but are typically poor at predicting responses to high-dimensional stimuli, such as natural images or movies. Recently however, it was shown how incorporating structured priors, e.g. for local and smooth receptive fields (RFs), can be used to scale up GPs to predict retinal neurons responses to static images [2]. However, it was unclear whether this approach could be used to predict neural responses to dynamic stimuli and movies, which have substantially higher dimensionality. Here we show that, by incorporating a recently proposed 'temporal relevance determination' (TRD) prior [3], which imposes a variable degree of smoothness as a function of time-lag, GPs can outperform a state-of-the art convolutional neural network (CNN) in predicting retinal responses to movies. Performance improvements were particularly marked when both models were trained on short recordings of less than 30 minutes. The GP had the additional advantage of outputting the uncertainty in its predictions (which could be used e.g. in optimal stimulus design [2]), and requiring relatively few 'transparent' prior assumptions about the network architecture. Moreover, it could be trained quickly, based on the responses of single neurons, while the CNN required responses from multiple retinal neurons to train the middle layers.

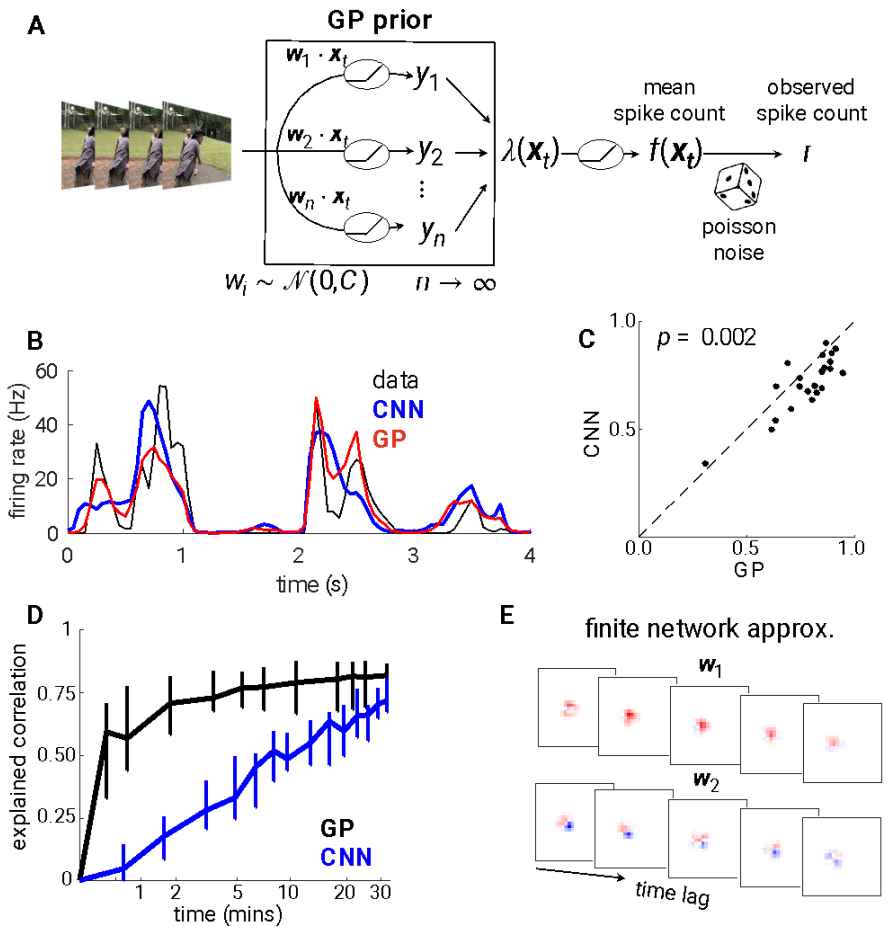

Fig 1: **(A)** The GP model is equivalent to a 2-layer neural network with infinite units in the middle layer. The model output is rectified & corrupted by Poisson noise to obtain the spike count. Spatio-temporal filters, $\boldsymbol{w}$, are *a priori* temporally/spatially smooth & local. **(B)** Mean firing rate of a single RGC in response to a dynamic, multi-scale, checkerboard stimulus (see text for details). The data (black) is plotted alongside predictions of the GP (red) and state of the art CNN model (blue). **(C)** The explained correlation for 22 cells (each cell is a dot) achieved by the GP versus the CNN model, when trained on the full 30 minute recording. **(D)** Performance of the CNN (blue) and GP (black) versus the length of training data. Solid lines are the median across cells, while vertical lines show the interquartile range. **(E)** Spatio-temporal filters of the 2 most active units, in a reduced 'finite network' approximation of the GP.

Methods

In our model, schematized in **Fig 1**, the observed spike count at each time point is drawn from a Poisson distribution with mean $f(x_t) = e^{A\lambda(x_t)+\lambda_0}$, where A and λ are scalar hyperparameters. We assume that the log-firing rate, $\lambda(x_t)$ was governed by a GP prior, with arc-cosine kernel $K_{arc-cos}(x_t, x_t')$ [4]. This kernel, is equivalent to a 2 layer neural network with infinite rectified linear units in the middle layer. In addition, we

assumed a zero-mean gaussian prior for the spatio-temporal filters in the first layer, with covariance $C = K_x \otimes K_t$, where K_x is the spatial prior favoring smooth & local RFs proposed by [2], and K_t is the TRD prior proposed recently by [3], which imposes a variable degree of smoothness as a function of time-lag. Plugging in this prior on the filters in the first layer results in a modified kernel, $K_{arc-cos}(C^{\frac{1}{2}} x_t, C^{\frac{1}{2}} x_t')$ ($C^{\frac{1}{2}}$ is the Cholesky decomposition of C). Inference and hyper-parameter learning was performed using an inducing point algorithm [5], to maximise a lower bound on the marginal log-likelihood.

We compared our model to a state-of-the-art deep CNN, proposed by McIntosh et al.. This details of this model are given in [1]. In brief, their model consists of 3 layers (2 convolutional and one fully connected), with time factorized in the first layer (i.e. only the first conv is non separable and has time). Hyperparameters (i.e. the L1 and L2 regularisation on the fully connected layer and the learning rate) were trained using our data-set of 22 retinal neurons.

To test our model, we used recorded spikes from 22 retinal ganglion cells (RGCs), in response to a 'multi-scale checkerboard' (MSC) stimulus, which consists of binary checkers, with checkers at multiple spatial scales (ranging from 8-? μm) and a refresh rate of 4Hz. (Note that, in contrast to the commonly used single scale checkerboard, a simple linear nonlinear model did poorly at predicting retinal responses to this stimulus). We fitted our models using spike counts in 50ms bins.

Results

We assessed the performance of the GP and CNN model in predicting the firing rate of each neuron to a repeated MSC stimuli, held out from the training dataset (**Fig 1B**). Note that, in contrast to the CNN, the GP had the added bonus of returning the uncertainty in its predictions (**Fig 1B** shaded red area), which could be useful in the for (i) testing hypotheses (because we can assess how reliable our predictions are) and/or (ii) closed-loop experiments [2].

We quantified model performance using the 'explained correlation' (the correlation between predicted/observed firing rate, normalized by reliability across trials). When trained on the full 30 minute recording, the GP scored significantly better than the CNN (**Fig 1C**, p=0.002, signed-rank test). Larger improvements were observed when we reduced the amount of training data (**Fig 1D**): e.g. with 5 mins recording, the median explained correlation was ~0.75 for the GP, and ~0.3 for the CNN. In addition, we note that the CNN was trained on data from all 22 cells, while the GP was trained on each cell individually.

As the GP is a black box model we cannot easily look inside to observe activations in the middle layer. To overcome this, we used a method proposed by [2] to derive a finite network approximation the GP (**Fig 1E**). This method involves replacing the arc-cosine kernel (with ∞ units in the middle layer) by a finite network approximation, and using a sparse prior to further remove units. While doing this resulted in a small reduction in performance (e.g. for the cell plotted in **Fig 1B,** the explained correlation dropped from 0.85 to 0.75 with a reduced network with 12 units), we can now see how individual units in the middle layer (with spatio-temporal filters plotted in **Fig 1E**) affect the output of the GP.

Summary

We propose a GP model for predicting neural responses to movies. We show that:

- With biologically inspired priors, we can scale up the GP to predict neural responses to movies.
- Compared to CNNs, the GP model requires few assumptions about network structure, and can thus could be readily applied to new data-sets, with little prior tuning.
- The GP model out-performs a state-of-the-art CNN in predicting retinal responses to dynamic movie stimuli. Improvements are largest when trained on little data.

- Goldin M. A., Virgili, S., Chalk M., (2023). Scalable gaussian process inference of neural responses to natural images *PNAS* 120 (34) e2301150120
- Duncker, L., Ruda, K. M., Field, G. D., & Pillow, J. W. (2023). Scalable Variational Inference for Low-Rank Spatiotemporal Receptive Fields. *Neural computation*, *35*(6), 995-1027.

- We derive a finite network approximation of the GP, to ‘look inside’ the model, and test for possible candidate mechanisms.