

Figure 2: Convolutional filter saliency over 150 epochs of SGD on CIFAR-10.

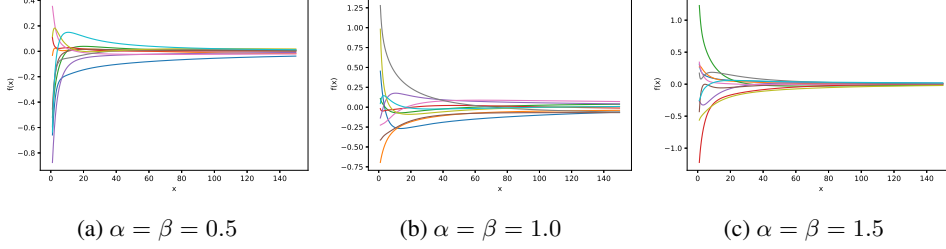


Figure 3: Function samples drawn from the exponential kernel.

## A DETAILS OF MOGP

### A.1 ON THE CHOICE OF THE “EXPONENTIAL KERNEL”

We justify our choice of the exponential kernel as a modeling mechanism by presenting visualizations of saliency measurements collected during training, and comparing these to samples drawn from the exponential kernel  $k_q(t, t') \triangleq \frac{\beta^\alpha}{(t+t'+\beta)^\alpha}$ , as shown in Figs. 2-3. Both the saliency and the function samples exhibit exponentially decaying behavior, which makes the exponential kernel a strong fit for modeling saliency evolution over time.

Furthermore we note that the exponential kernel was used to great effect in Swersky et al. (2014) with respect to modeling loss curves as a function of epochs. Loss curves also exhibit asymptotic behavior, similar to saliency measurement curves, thus providing evidence for the exponential kernel being an apt fit for our task.

### A.2 PREDICTIVE DISTRIBUTION OF THE SALIENCY

Given a vector of observed saliency  $\tilde{\mathbf{s}}_{1:t}$ , the MOGP regression model can provide a Gaussian predictive distribution  $p(\mathbf{s}_{t'}|\tilde{\mathbf{s}}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_{t'|1:t}, \mathbf{K}_{t'|1:t})$  for any future saliency  $\mathbf{s}_{t'}$  with the following posterior mean vector and covariance matrix:  $\boldsymbol{\mu}_{t'|1:t} \triangleq \boldsymbol{\mu}_{t':t'} + \mathbf{K}_{[t't]} \mathbf{K}_{1:t}^{-1} (\tilde{\mathbf{s}}_{1:t} - \boldsymbol{\mu}_{1:t})$ ,  $\mathbf{K}_{t'|1:t} \triangleq \mathbf{K}_{t':t'} - \mathbf{K}_{[t't]} \mathbf{K}_{1:t}^{-1} \mathbf{K}_{[t't]}^\top$  where  $\mathbf{K}_{[t't]} \triangleq [\text{cov}[s_{t'}^a, s_{\tau'}^{a'}]]_{\tau=1, \dots, t}^{a, a'=1, \dots, M}$ . Then, the  $a$ -th element  $\mu_{t'|1:t}^a$  of  $\boldsymbol{\mu}_{t'|1:t}$  is the predictive mean of the saliency  $s_{t'}^a$ . And the  $[a, a']$ -th element of  $\mathbf{K}_{[t't]}$  denoted as  $\sigma_{t'|1:t}^{aa'}$  is the predictive (co)variance between the saliency  $s_{t'}^a$  and  $s_{t'}^{a'}$ .

## B SUBMODULARITY OF $\mathbb{E}[\hat{\rho}_T]$

In (7), the problem of choosing  $\mathbf{m}$  from  $\{0, 1\}^M$  can be considered as selecting a subset  $A$  of indexes from  $\{1, \dots, M\}$  such that  $m_t^a = 1$  for  $a \in A$ , and  $m_t^a = 0$  otherwise. Therefore,  $P(\mathbf{m}) \triangleq \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})}[\hat{\rho}_T(\mathbf{m}, B_s)]$  can be considered as a set function which we will show to be submodular. To keep notation consistency, we will remain using  $P(\mathbf{m})$  instead of representing it as a function of the index subset  $A$ .

**Lemma 2 (Submodularity).** *Let  $\mathbf{m}', \mathbf{m}'' \in \{0, 1\}^M$ , and  $e^{(a)}$  be arbitrary  $M$ -dimensional one hot vector with  $1 \leq a \leq M$ . We have  $P(\mathbf{m}' \vee e^{(a)}) - P(\mathbf{m}') \geq P(\mathbf{m}'' \vee e^{(a)}) - P(\mathbf{m}'')$  for any  $\mathbf{m}' \preceq \mathbf{m}''$ ,  $\mathbf{m}' \wedge e^{(a)} = 0$ , and  $\mathbf{m}'' \wedge e^{(a)} = 0$ .*

*Proof.* According to (4),

$$\mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})}[\hat{\rho}_T(\mathbf{m}, B_s)] = \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} \left[ \max_{\mathbf{m}_T} [\mathbf{m}_T \cdot \tilde{\mathbf{s}}_T, \text{ s.t. } \|\mathbf{m}_T\|_0 \leq B_s, \mathbf{m}_T \preceq \mathbf{m}] \right]$$

Let  $\alpha(\mathbf{m}) \triangleq \arg \max_{\mathbf{m}_T} [\mathbf{m}_T \cdot \tilde{\mathbf{s}}_T, \text{ s.t. } \|\mathbf{m}_T\|_0 \leq B_s, \mathbf{m}_T \preceq \mathbf{m}]$  return the optimized mask  $\mathbf{m}_T$  given any  $\mathbf{m}$ ,  $\Lambda_{\mathbf{m}} \triangleq \min(\alpha(\mathbf{m}) \odot \mathbf{s}_T)$  be the minimal saliency of the network elements selected at iteration  $T$  for  $P(\mathbf{m})$ . Then, we have

$$\begin{aligned} P(\mathbf{m} \vee e^{(a)}) &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\hat{\rho}_T(\mathbf{m} \vee e^{(a)}, B_s)] \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\hat{\rho}_T(\mathbf{m}, B_s) - \Lambda_{\mathbf{m}} + \max(s_T^a, \Lambda_{\mathbf{m}})] \end{aligned}$$

The second equality is due to the fact that the network element  $v_T^a$  would only replace the lowest included element in  $\mathbf{m}_T$  in order to maximize the objective. Then,

$$\begin{aligned} &P(\mathbf{m} \vee e^{(a)}) - P(\mathbf{m}) \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\hat{\rho}_T(\mathbf{m}, B_s) - \Lambda_{\mathbf{m}} + \max(s_T^a, \Lambda_{\mathbf{m}})] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\hat{\rho}_T(\mathbf{m}, B_s)] \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [-\Lambda_{\mathbf{m}} + \max(s_T^a, \Lambda_{\mathbf{m}})] \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^a - \Lambda_{\mathbf{m}}, 0)] \end{aligned} \tag{9}$$

Given  $\mathbf{m}' \preceq \mathbf{m}''$ , we have  $\Lambda_{\mathbf{m}'} \leq \Lambda_{\mathbf{m}''}$  since  $\mathbf{m}_T \preceq \mathbf{m}$  in  $\alpha(\mathbf{m}')$  is a tighter constraint than that in  $\alpha(\mathbf{m}'')$ . Consequently, we can get  $s_T^a - \Lambda_{\mathbf{m}'} \geq s_T^a - \Lambda_{\mathbf{m}''}$ , and thus

$$[P(\mathbf{m}' \vee e^{(a)}) - P(\mathbf{m}')] \geq [P(\mathbf{m}'' \vee e^{(a)}) - P(\mathbf{m}'')].$$

□

## C PROOF OF LEMMA 1

We restate Lemma 1 for clarity.

**Lemma 1.** Let  $\mathbf{e}^{(i)}$  be an  $M$ -dimensional one-hot vectors with the  $i$ -th element be 1.  $\forall 1 \leq a, b \leq M$ ;  $\mathbf{m} \in \{0, 1\}^M$  s.t.  $\mathbf{m} \wedge (\mathbf{e}^{(a)} \vee \mathbf{e}^{(b)}) = 0$ . Given a vector of observed saliency  $\tilde{\mathbf{s}}_{1:t}$ , if  $\mu_{T|1:t}^a \geq \mu_{T|1:t}^b$  and  $\mu_{T|1:t}^a \geq 0$ , then  $\mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(b)})] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(a)})] \leq \mu_{T|1:t}^b \Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) + \theta \phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right)$  where  $\Phi$  and  $\phi$  are standard normal CDF and PDF respectively, and  $\theta \triangleq \sqrt{\sigma_{T|1:t}^{aa} + \sigma_{T|1:t}^{bb} - 2\sigma_{T|1:t}^{ab}}$ . In particular, if  $\frac{\mu_{T|1:t}^a - \mu_{T|1:t}^b}{\theta} \gg 1$ , then  $\mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(b)})] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(a)})] \leq \epsilon$ .

To prove this Lemma, we prove the following first:

**Lemma 3.**  $\mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(b)})] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(a)})] \leq \mathbb{E}[\max(s_T^b - s_T^a, 0)]$ .

*Proof.* Due to (9), we have

$$\begin{aligned} &\mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(b)})] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\rho_T(\mathbf{m} \vee \mathbf{e}^{(a)})] \\ &= P(\mathbf{m} \vee \mathbf{e}^{(b)}) - P(\mathbf{m}) - (P(\mathbf{m} \vee \mathbf{e}^{(a)}) - P(\mathbf{m})) \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^b - \Lambda_{\mathbf{m}}, 0)] - \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^a - \Lambda_{\mathbf{m}}, 0)] \\ &= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^b - \Lambda_{\mathbf{m}}, 0) - \max(s_T^a - \Lambda_{\mathbf{m}}, 0)] \end{aligned} \tag{10}$$

$$= \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^b - s_T^a, \Lambda_{\mathbf{m}} - s_T^a) - \max(0, \Lambda_{\mathbf{m}} - s_T^a)] \tag{11}$$

$$\leq \mathbb{E}_{p(\mathbf{s}_T|\tilde{\mathbf{s}}_{1:t})} [\max(s_T^b - s_T^a, 0)] \tag{12}$$

The equality (11) is achieved by adding  $\Lambda_{\mathbf{m}} - s_T^a$  in each term of the two max functions in (10). The inequality (12) can be proved by considering the following two cases:

Table 4: Comparing log likelihood of test data for Independent GPs (GP) vs. MOGP with n latent functions (n-MOGP) on collected saliency measurements from CIFAR-10 training.

	Small dataset			Medium dataset			Large dataset		
	Lyr 1	Lyr 2	Lyr 3	Lyr 1	Lyr 2	Lyr 3	Lyr 1	Lyr 2	Lyr 3
GP	-1.19	-1.08	-1.07e5	-0.96	-0.93	-2.47	-0.49	-0.48	-1.33
4-MOGP	-1.15	-0.89	-2.44	-0.91	-0.80	-2.20	-0.38	-0.39	-1.25
8-MOGP	-1.09	-0.86	-2.38	-0.84	-0.78	-2.16	-0.32	-0.35	-1.20
18-MOGP	-0.97	<b>-0.80</b>	-2.33	-0.89	-0.76	-2.13	-0.31	-0.35	-1.20
32-MOGP	<b>-0.96</b>	-0.81	<b>-2.32</b>	<b>-0.79</b>	<b>-0.74</b>	<b>-2.13</b>	<b>-0.31</b>	<b>-0.34</b>	<b>-1.20</b>

If  $\Lambda_{\mathbf{m}} - s_T^a \geq 0$ , then

$$\begin{aligned}
& \max(s_T^b - s_T^a, \Lambda_{\mathbf{m}} - s_T^a) - \max(0, \Lambda_{\mathbf{m}} - s_T^a) \\
&= \max(s_T^b - s_T^a, \Lambda_{\mathbf{m}} - s_T^a) - (\Lambda_{\mathbf{m}} - s_T^a) \\
&= \max(s_T^b - s_T^a - (\Lambda_{\mathbf{m}} - s_T^a), 0) \\
&\leq \max(s_T^b - s_T^a, 0).
\end{aligned}$$

If  $\Lambda_{\mathbf{m}} - s_T^a < 0$ , then

$$\begin{aligned}
& \max(s_T^b - s_T^a, \Lambda_{\mathbf{m}} - s_T^a) - \max(0, \Lambda_{\mathbf{m}} - s_T^a) \\
&= \max(s_T^b - s_T^a, \Lambda_{\mathbf{m}} - s_T^a) \\
&\leq \max(s_T^b - s_T^a, 0).
\end{aligned}$$

□

Next we utilize a well known bound regarding the maximum of two Gaussian random variables (Nadarajah & Kotz, 2008), which we restate:

**Lemma 4.** *Let  $s^a, s^b$  be Gaussian random variables with means  $\mu^a, \mu^b$  and standard deviations  $\sigma^a, \sigma^b$ , then  $\mathbb{E}[\max(s^a, s^b)] \leq \mu^a \Phi\left(\frac{\mu^b - \mu^a}{\theta}\right) + \mu^b \Phi\left(\frac{\mu^b - \mu^a}{\theta}\right) + \theta \phi\left(\frac{\mu^b - \mu^a}{\theta}\right)$  where  $\theta \triangleq \sqrt{[\sigma^b]^2 + [\sigma^a]^2 - 2\text{cov}(s^b, s^a)}$  and  $\Phi, \phi$  are standard normal CDF and PDF respectively.*

Then,

$$\begin{aligned}
& \mathbb{E}_{p(s_T|\bar{s}_{1:t})}[\max(s_T^b - s_T^a, 0)] \\
&= \mathbb{E}_{p(s_T|\bar{s}_{1:t})}[\max(s_T^b, s_T^a)] - \mathbb{E}_{p(s_T|\bar{s}_{1:t})}[s_T^a] \\
&\leq (\mu_{T|1:t}^b + \mu_{T|1:t}^a) \Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) + \theta \phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) - \mu_{T|1:t}^a \\
&= \mu_{T|1:t}^b \Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) + \theta \phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) + \mu_{T|1:t}^a \left( \Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) - 1 \right) \\
&\leq \mu_{T|1:t}^b \Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) + \theta \phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right)
\end{aligned}$$

The first inequality follows from Lemma 4. The second inequality is due to  $\Phi\left(\frac{\mu_{T|1:t}^b - \mu_{T|1:t}^a}{\theta}\right) \leq 1$  and  $\mu_{T|1:t}^a \geq 0$ .

## More Experimental Results and Experimental Details

### C.1 GP VS. MOGP LOG-LIKELIHOOD ON CIFAR-10 DATASET

Table 4 presents the results of the experiment in Section 4.1 for the CIFAR-10 dataset.

## C.2 DATA PREPROCESSING

We followed the same data preprocessing procedure for both our small scale and ImageNet experiments. To standardize the saliency measurements for a training dataset  $\tilde{s}_{1:t}$  in our modeling experiments we clip them between 0 and an upper bound computed as follows:  $ub \triangleq \text{percentile}(\tilde{s}_{1:t}, 95) \times 1.3$ . This procedure removes outliers. We used 1.3 as a multiplier, as this upper bound is used to transform test dataset as well, which may have higher saliency evaluations.

After clipping the training data, we perform a trend check for each element  $v^a$  by fitting a Linear Regression model to the data  $\tilde{s}_{1:t}^a$ . For  $\tilde{s}_{1:t}^a$  with an increasing trend (i.e., the linear regression model has positive slope) we perform the transformation  $\tilde{s}_{1:t}^a = ub - \tilde{s}_{1:t}^a$ . The reasoning behind this is that the exponential kernel strongly prefers *decaying* curves. After this preprocessing, we scale up the saliency measurements to a  $[0, 10]$  range:  $\tilde{s}_{1:t} = \tilde{s}_{1:t} \times 10$ . We found that without scaling to larger values, log-likelihood of our models demonstrated extremely high positive values due to small values of unscaled saliency measurements.

We transform the test data in our modeling experiments  $\tilde{s}_{t+1:T}$  with the same procedure using the same  $ub$  and per-element  $v^a$  regression models as computed by the training data. We measure log-likelihood after this transformation for both the test dataset in our small scale experiments.

During the BEP Algorithm, the same steps are followed, however we inverse the trend check transformation ( $\tilde{s}_{1:t}^a = ub - \tilde{s}_{1:t}^a$ ) on the predicted MOGP distribution of  $s_T$  prior to sampling for estimation of  $\Delta(\cdot)$ .

## C.3 EXPERIMENTAL DETAILS

To train our CIFAR-10 and CIFAR-100 models we used an Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.001. The learning rate used an exponential decay of  $k = 0.985$ , and a batch size of 32 was used. Training was paused three times evenly spaced per epoch. During this pause, we collected saliency measurements using 40% of the training dataset. This instrumentation subset was randomly select from the training dataset at initialization, and remained constant throughout the training procedure. We performed data preprocessing of saliency evaluations into a standardized  $[0, 10]$  range.<sup>15</sup> We used (3) to measure saliency of neurons/convolutional filters. For the convolutional layers we used 12 latent MOGP functions. For the dense layer we used 4 latent MOGP functions.

For our ResNet-50 model we used an SGD with Momentum optimizer with an initial learning rate of 0.1. The learning rate was divided by ten at  $t = [30, 60, 80]$  epochs. We collected saliency data every 5 iterations of SGD, and averaged them into buckets corresponding to 625 iterations of SGD to form our dataset. We used a minimum of 4 latent functions per MOGP, however this was dynamically increased if the model couldn't fit the data up to a maximum of 15.

We sampled 10K points from our MOGP model to estimate  $\Delta(\cdot)$  for CIFAR-10/CIFAR-100. For ResNet we sampled 15K points. We repeated experiments 5 times for reporting accuracy on CIFAR-10/CIFAR-100.

## C.4 PRUNING ON RESNET

ResNet architecture is composed of a sequence of residual units:  $Z_\ell \triangleq \mathcal{F}(\mathbf{P}_{\ell-1}) + \mathbf{P}_{\ell-1}$ , where  $\mathbf{P}_{\ell-1}$  is the output of the previous residual unit  $Z_{\ell-1}$  and '+' denotes elementwise addition. Internally,  $\mathcal{F}$  is typically implemented as three stacked convolutional layers:  $\mathcal{F}(\mathbf{P}_{\ell-1}) \triangleq [z_{\ell_3} \circ z_{\ell_2} \circ z_{\ell_1}](\mathbf{P}_{\ell-1})$  where  $z_{\ell_1}, z_{\ell_2}, z_{\ell_3}$  are convolutional layers. Within this setting we consider convolutional filter pruning. Although  $z_{\ell_1}, z_{\ell_2}$  may be pruned using the procedure described earlier. Pruning  $z_{\ell_3}$  requires a different procedure. Due to the direct addition of  $\mathbf{P}_{\ell-1}$  to  $\mathcal{F}(\mathbf{P}_{\ell-1})$ , the output dimensions of  $Z_{\ell-1}$  and  $z_{\ell_3}$  must match exactly. Thus a ResNet architecture consists of sequences of residual units of length  $B$  with matching input/output dimensions:  $\zeta \triangleq [Z_\ell]_{\ell=1, \dots, B}$ , s.t.  $\dim(\mathbf{P}_1) = \dim(\mathbf{P}_2) = \dots = \dim(\mathbf{P}_B)$ . We propose *group pruning* of layers  $[z_{\ell_3}]_{\ell=1, \dots, B}$  where filters are removed from all  $z_{\ell_3}$  in a residual unit sequence in tandem. We define  $s([\zeta, c]) \triangleq \sum_{\ell=1}^B s([z_{\ell_3}, c])$ , where  $s(\cdot)$  is defined for convolutional layers as in (3). To prune the channel  $c$  from  $\zeta$ , we prune it from each layer

<sup>15</sup>Generally, saliency evaluations are relatively small ( $\leq 0.01$ ), which leads to poor fitting models or positive log-likelihood. Precise details of our data preprocessing is in Appendix C.2.

in  $[z_{\ell 3}]_{\ell=1,\dots,B}$ . Typically we pruned sequence channels less aggressively than convolutional filters as these channels feed into several convolutional layers.