Function-space Parameterization of Neural Networks for Sequential Learning Aidan Scannell*, Riccardo Mereu*, Paul Edmund Scan Chang, Ella Tamir, Joni Pajarinen, Arno Solin For project website+code *Equal contribution









Motivation: Combine Benefits of Neural Networks and Gaussian Processes





Uncertainty Quantification Neural networks lack uncertainty estimates





Uncertainty Quantification Neural networks lack uncertainty estimates

Neural network







Uncertainty Quantification Neural networks lack uncertainty estimates

Neural network





Gaussian process





Image Inputs Neural networks scale to high-dimensional inputs





Image Inputs Neural networks scale to high-dimensional inputs

Neural network

airplane	and a	X	-	X	*	*	2	-4	-	8
automobile					-	Tes			-	*
bird	Can and a start of the start of	ſ	2			-	1	Y	- Har	4
cat			4	64		1	Z.	Å.	No.	1
deer	1	48	X	RA		Y	Ŷ	1	-	
dog	科	1	-	B .	1	(A)	9	13	A	1¢
frog	-7	19	1		2			5		5
horse	-	-	1	2	1	107AB	-	20		N.
ship			dista	-	<u>M</u>	-	2	NET	1	
truck										ALL.





Image Inputs Neural networks scale to high-dimensional inputs

Neural network

airplane	and a	X	-	X	*	*	2	-4	-	8
automobile					-	Tes			-	*
bird	Se la	ſ	2			-	1	Y	- Har	4
cat			4	64		1	Z.	Å.	No.	1
deer	1	48	X	RA		Y	Ŷ	1	-	
dog	科	1	-	B .	1	(A)	9	13	A	1¢
frog	-7	19	1		2			5		5
horse	-	-	1	2	1	107AB	-	24		N.
ship			dista	-	<u>M</u>	-	2	NET	1	
truck										ALL.



Gaussian process



$$K(\vec{x}, \vec{l}^i) = e^{-\frac{\|\vec{x} - \vec{l}^i\|^2}{2\sigma^2}}$$

Image source: http://www.cs.toronto.edu/~duvenaud/cookbook/index.htm

Large Data Sets Neural networks scale to large data sets





Large Data Sets Neural networks scale to large data sets

Neural network







Large Data Sets Neural networks scale to large data sets

Neural network





Gaussian process





Incorporating New Data Neural networks require costly retraining





Incorporating New Data Neural networks require costly retraining

Neural network







Incorporating New Data Neural networks require costly retraining

Neural network





Gaussian process





Combines benefits of neural networks and Gaussian processes







Train Data

+

++ + ++ +







Weight-space NN training



Train Data

+

╋

++ _+++ +

 w_1

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{x}_i)}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \frac{\mathcal{R}(\mathbf{x}_i)}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \frac{\mathcal{R}(\mathbf{x}_i)}{-\log$$







Weight-space NN training



 w_1

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{x}_i)}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \frac{-\log q}{-\log q}$$

Neural network



Weight-space NN training



 w_1

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{w})}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \underbrace{\mathscr{R}(\mathbf{w})}{-\log p(\mathbf{w})}$$
FCAI



Weight-space NN training



Linearise to get



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{w})}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \underbrace{\mathscr{R}(\mathbf{w})}{-\log p(\mathbf{w})}$$
FCAI

kernel formulation





Weight-space NN training



Linearise to get



$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{w})}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \underbrace{\mathscr{R}(\mathbf{w})}{-\log p(\mathbf{w})}$$
FCAI





Weight-space NN training



Linearise to get



FCAI

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \mathscr{R}(\mathbf{w})}{-\log p(y_i | f_{\mathbf{w}}(\mathbf{x}_i))} + \underbrace{\mathscr{R}(\mathbf{w})}_{-\log p(\mathbf{w})}$$





$p(\mathbf{f} \mid \mathcal{D}) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mathbf{f^*}, \mathbf{S_f})$

Function-space Laplace approx. fcai.fi



$p(\mathbf{f} \mid \mathscr{D}) \approx q(\mathbf{f}) = \mathscr{N}(\mathbf{f} \mid \mathbf{f}^*, \mathbf{S}_{\mathbf{f}})$ Laplace approx.



Function-space





$$p(\mathbf{f} \mid \mathscr{D}) \approx q(\mathbf{f}) = \mathscr{N}(\mathbf{f} \mid \mathbf{f}^*, \mathbf{S}_{\mathbf{f}})$$
 La

. GP predictive posterior is computationally expensive ullet



Function-space place approx.





$$p(\mathbf{f} \mid \mathscr{D}) \approx q(\mathbf{f}) = \mathscr{N}(\mathbf{f} \mid \mathbf{f}^*, \mathbf{S}_{\mathbf{f}})$$
 Lap

- <u>•</u> GP predictive posterior is computationally expensive
- We sparsify the GP using a dual parameterization



Function-space place approx.



















Uncertainty estimates 🔽



















Incorporate new data fast 🔽













Number of inducing points

Number of inducing points

Number of inducing points

FCAI

Higher complexity

FCAI

Number of inducing points

Higher complexity

FCAI

Number of inducing points

Higher complexity

SFR for Sequential Learning

Input, *x* **layers. Left. Predictions from the principal of a** the training data with the help of a and uncertainty, and (right) makes it representation for continual learning

 $\mathbf{2}$

ring the contributi ertainty. Crucially etwork, with the b otorious cubic con nection between th d GPs [67]. We re rived from a traine orm dual conditions to retrain the mo

0

2

SFR, a new appro: F-MNIST GP Subset Updating the model post-

☆ Memory points — Decision boundary / Predictive probability

Input, *x* **layers. Left Predictions** from **the raining** the training data with the help of a and uncertainty, and (right) makes it representation for continual learning

 $\mathbf{2}$

ring the contributi ertainty. Crucially etwork, with the b otorious cubic con nection between th d GPs [67]. We re rived from a traine orm dual conditic g to retrain the mo

0

SFR, a new appro: F-MNIST that, d GP Subset updating the model post-

☆ Memory points — Decision boundary / Predictive probability

Input, *x* layers. Left Predictions from the rning the training data with the help of a and uncertainty, and (right) makes it representation for continual learning

 $\mathbf{2}$

ring the contributi ertainty. Crucially etwork, with the b otorious cubic con nection between the al GPs [67]. We re rived from a traine orm dual conditic g to retrain the mo

0

SFR, a new approa **F-MNIST** that, de - SFR - GP Subset updating the model post-(ii) Wa provide extensive

 \cancel{K} Memory points — Decision boundary / Predictive probability

Input, *x* layers. Left Predictions from the rning the training data with the help of a and uncertainty, and (right) makes it representation for continual learning

2

ring the contributi ertainty. Crucially etwork, with the b otorious cubic con nection between the al GPs [67]. We re rived from a traine orm dual conditic g to retrain the mo

0

SFR, a new approa **F-MNIST** that, de - SFR - GP Subset updating the model post-(ii) Wa provide extensive

 \cancel{K} Memory points — Decision boundary / Predictive probability

Input, *x* layers. Left Predictions from the principal of a the training data with the help of a and uncertainty, and (right) makes it representation for continual learning

2

ring the contributi ertainty. Crucially etwork, with the b otorious cubic con nection between the al GPs [67]. We re rived from a traine orm dual conditic g to retrain the mo

SFR, a new approa

- SFR

- GP Subset

that, de

F-MNIST

0

🗕 Data	☆ Memory points	— Decision bo	undary 🦰 / 🗖 F	Predictive probability
Method	S-MNIST (SH) 40 pts./task	S-MNIST (SH) 200 pts./task	S-FMNIST (SH) 200 pts./task) P-MNIST (SH) 200 pts./task
DER	$85.26{\scriptstyle \pm 0.54}$	$92.13{\scriptstyle \pm 0.45}$	$82.03 \scriptstyle \pm 0.57$	$93.08{\scriptstyle \pm 0.11}$
FROMP	$75.21{\scriptstyle \pm 2.05}$	$89.54{\scriptstyle \pm 0.72}$	$78.83{\scriptstyle \pm 0.46}$	$94.90{\scriptstyle \pm 0.04}$
S-FSVI	$84.51{\scriptstyle\pm1.30}$	$92.87{\scriptstyle \pm 0.14}$	$77.54{\scriptstyle \pm 0.40}$	$95.76{\scriptstyle \pm 0.02}$
SFR (Ours)	$89.22 {\scriptstyle \pm 0.76}$	$94.19 \scriptstyle \pm 0.26$	$81.96{\scriptstyle \pm 0.24}$	$95.58{\scriptstyle \pm 0.08}$

updating the model postii) Wa provida avtanciva

Reinforcement Learning SFR's uncertainty can guide exploration in model-based RL

Reinforcement Learning SFR's uncertainty can guide exploration in model-based RL

Cartpole swingup setup

Reinforcement Learning SFR's uncertainty can guide exploration in model-based RL

SFR (Ours)	GP	NN

Uncertainty estimates

SFR (Ours)	GP	NN
		X

Uncertainty estimates

Image inputs

SFR (Ours)	GP	NN
		×
	X	

Unc	ertainty estimates	
Imaç	ge inputs	
Larg	e data	

SFR (Ours)	GP	NN
		X
	X	
	×	

Uncertainty estimates	
Image inputs	
Large data	
Incorporate new data fast	

SFR (Ours)	GP	NN
		×
	X	
	X	
		X

Project Website

https://aaltoml.github.io/sfr/

Aidan Scannell*, Riccardo Mereu*, Paul Edmund Chang, Ella Tamir, Joni Pajarinen, Arno Solin

*Equal contribution

