National Nuclear Security Administration

Incorporating Prior Knowledge in Deep Learning Models

David Carlson

Duke University

February 8th, 2023

Duke UNIVERSITY

Consortium for ENABLING TECHNOLOGIES & INNOVATION

Introduction and Motivation

- Machine Learning and Artificial Intelligence (ML/AI) techniques are transformative but *data-hungry*
- We often have significant prior knowledge about data trends (e.g., seasonality in remote sensing)
- Incorporating prior knowledge is:
  - Challenging in deep learning
  - Routine in probabilistic models (e.g., Gaussian processes)

- Can we exploit relationships between deep neural networks and Gaussian processes to use prior information?

Mission Relevance

- Reducing sample complexity can improve nuclear monitoring from space-based sensors through more efficient data analysis
- AI models can be trained with fewer resources, leading to faster and more accurate detection of potential nuclear proliferation activities
- Additionally, reducing sample complexity can also improve the ability of these models to identify patterns and anomalies in the data, which is crucial for detecting potential nonproliferation activities

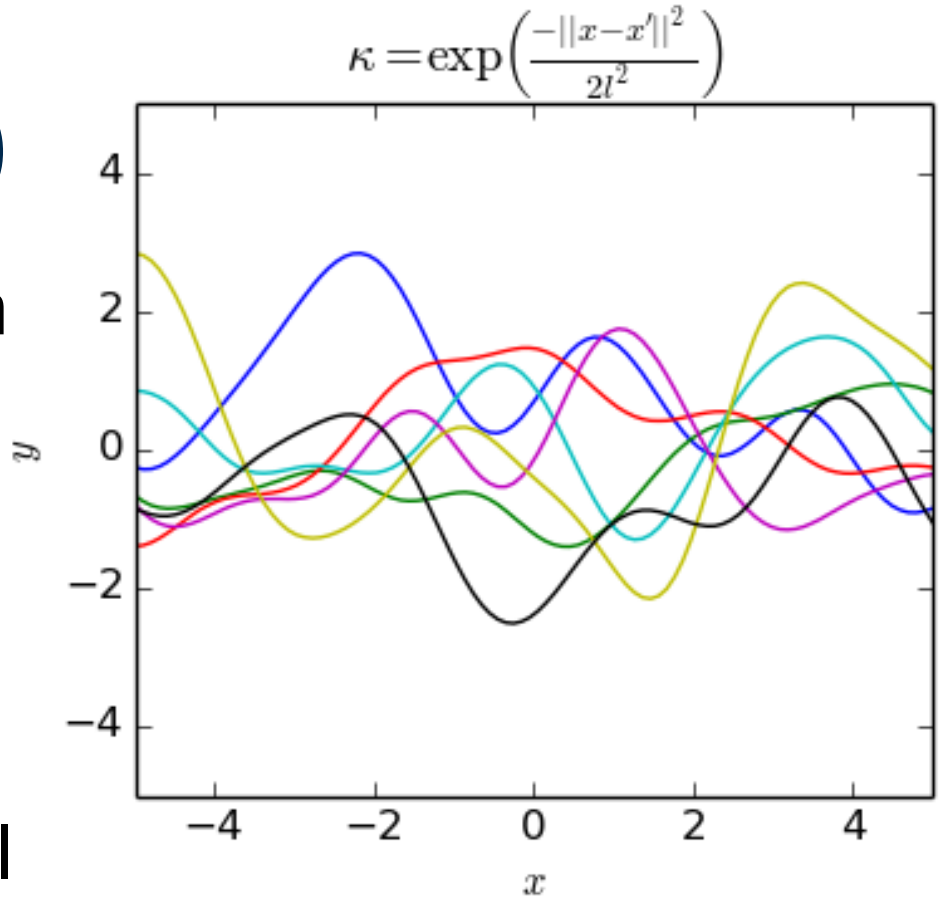
Composite Kernels for Gaussian Processes (GPs)

- Gaussian processes are probabilistic models for learning smooth functions from a mean function $\mu(\cdot)$ and covariance kernel $k(\cdot,\cdot)$:

$$f \sim GP(\mu, k)$$

- The kernel function encodes relationships between data points. Suppose that we have data made up of two modalities $x = \left[x^{(1)}, x^{(2)}\right]$, we can use a composite kernel

$$k(x_i, x_j) = k_1\left(x_i^{(1)}, x_j^{(1)}\right) \times k_2(x_i^{(2)}, x_j^{(2)})$$

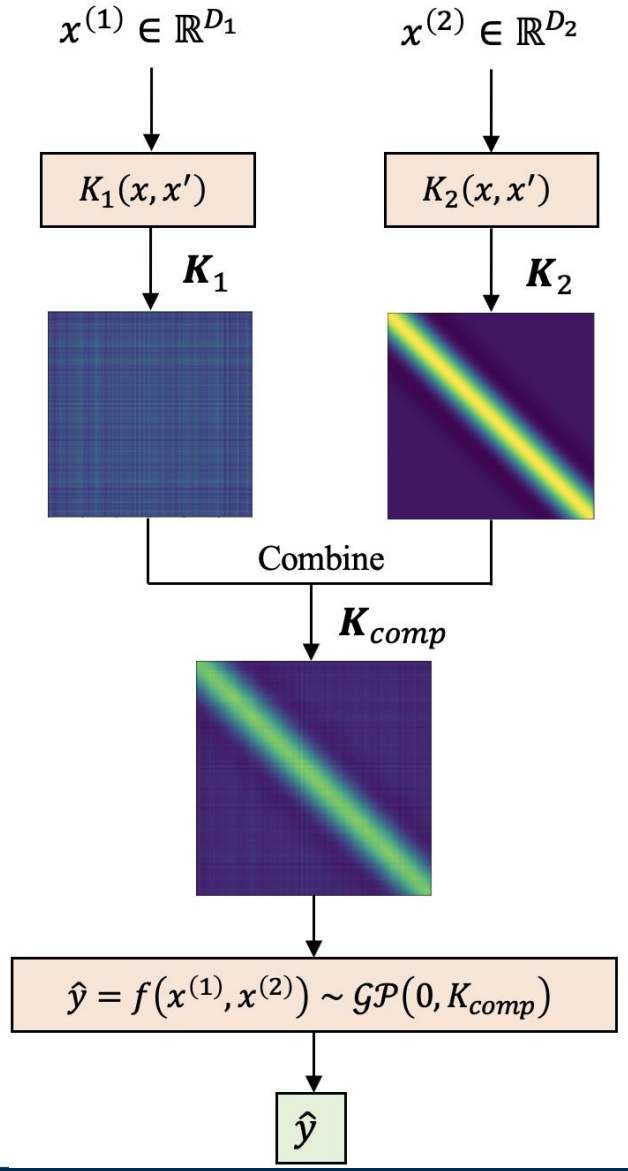


Cdipaolo96, CC BY-SA 4.0 via Wikimedia Commons

Composite Kernels are common for spatiotemporal data

- Can consider first modality as covariates and second modality as time/space
- Easy to enforce periodicities or spatial smoothing by choosing k_2 to encode desired relationship
- For example, seasonality can be encoded by a periodic kernel:

$$k_2 = \exp(-2\sin^2(d/2)/\ell^2)$$

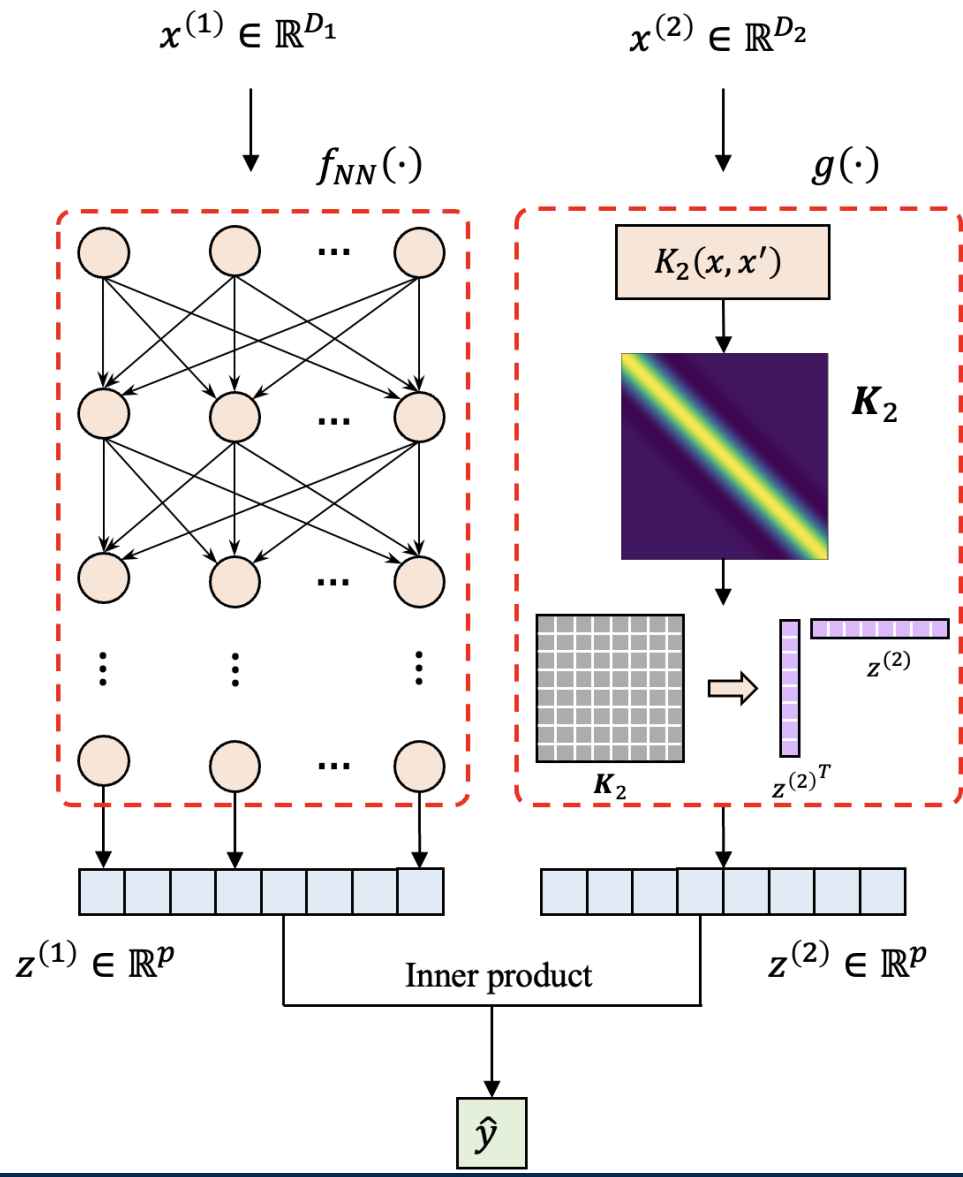


Neural Networks Approximate GPs

- A randomly initialized Neural Network approaches a GP with an implicit kernel k_{NN} (with a few assumptions):

$$f_{NN}(x) \rightarrow GP(0, k_{NN})$$

- We propose to use a neural network to handle the covariate data (e.g., satellite image or sensor data) while explicitly defining the second relationship



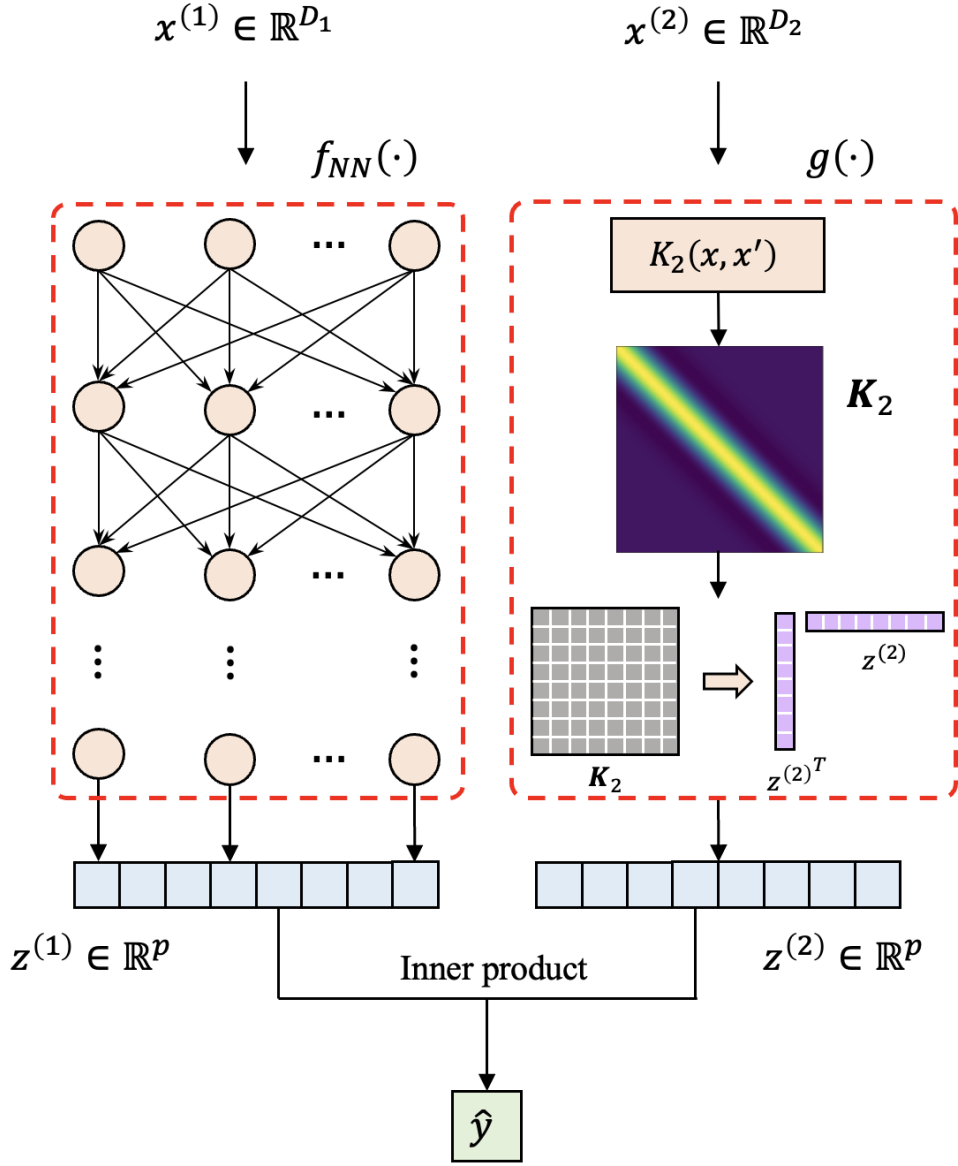
Approximates a Composite Kernel GP

- Theorem 4.1 of Jiang et al proves that this approach approximates a composite kernel if we can define the mapping:

$$z^{(2)} = g\left(x^{(2)}\right) \quad s.t.$$

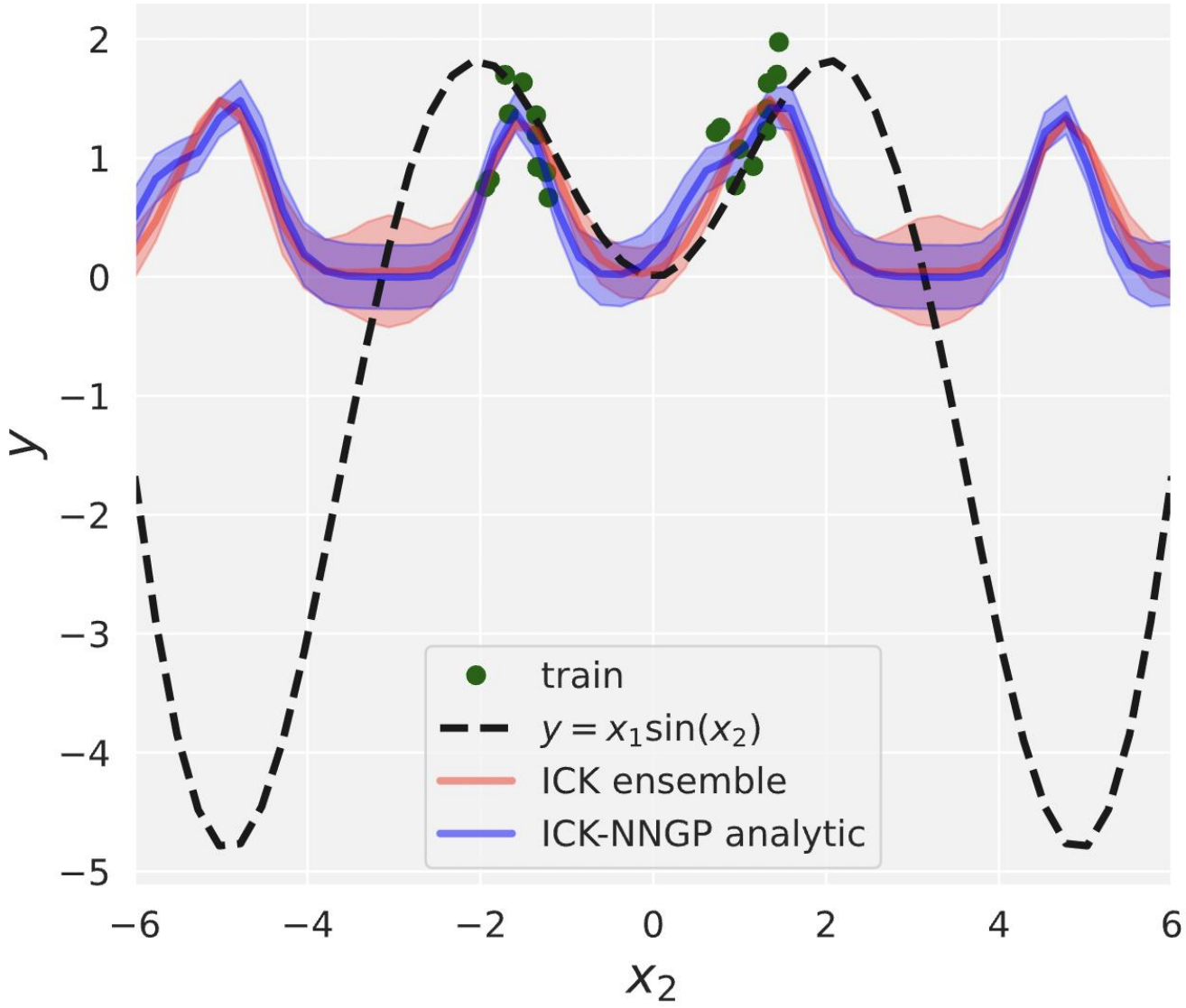
$$k^2\left(x_i^{(2)}, x_j^{(2)}\right) \simeq g\left(x_i^{(2)}\right)^T g\left(x_i^{(2)}\right)$$

- We can construct such a mapping from a Nyström approximation and Cholesky Decomposition in $O(p^3)$ time



Empirically matches theoretical posterior

- Can approximate a posterior by ensembling different neural network initializations
- For specific neural networks, the implicit kernel is known and the posterior can be analytically calculated.



Empirically Improves Remote Sensing Tasks and Seasonal Predictions

Table 1. Correlation and error statistics of ICK*y* and other joint deep models with both convolutional and attention-based architectures on the $PM_{2.5}$ forecasting task. "S." denotes seasonal variants.

	R_{Spear}	RMSE	MAE	MSLL
CNN-RF	0.00	194.63	185.83	-
ViT-RF	0.07	190.82	181.63	-
S. CNN-RF	0.62	53.36	39.38	96.77
S. ViT-RF	0.66	56.45	41.73	14.69
S. Deep-ViT-RF	0.65	56.36	42.46	17.63
S. MAE-ViT-RF	0.67	53.87	40.78	31.09
CNN-ICK*y*	0.62	53.46	39.76	**10.92**
ViT-ICK*y*	**0.68**	56.56	41.41	12208
DeepViT-ICK*y*	0.66	**52.41**	**35.93**	38220

Table 2. Prediction error of actual worker productivity on the test data set with ICK*y* and other benchmark models (MLPs and NPs)

	MSE $\downarrow$ $(*10^{-3})$	MAE $\downarrow$ $(*10^{-2})$
MLP	20.16 ± 1.26	9.93 ± 0.36
Cyclic MLP	20.97 ± 1.98	10.16 ± 0.77
GNP	57.25 ± 4.31	19.39 ± 0.94
AGNP	43.11 ± 5.95	14.38 ± 0.88
ICK*y*, $T = 2$	3.43 ± 1.42	4.85 ± 1.00
ICK*y*, $T = 7$	0.44 ± 0.13	1.43 ± 0.15
ICK*y*, $T = 30$	$\mathbf{0.31 \pm 0.09}$	$\mathbf{1.17 \pm 0.14}$

Conclusion

- We can encode prior knowledge into deep neural networks by defining a kernel on part of the data
  - Straightforward to incorporate known spatial or temporal relationships
- Algorithmic framework is theoretically proven and empirically robust
- Can improve predictive performance or reduce data necessary for acceptable performance

ETI Impact

- ETI funding and conversations have helped support and clarify this research vision
  - Students went to national labs this past summer
- We are very interested in applying this technology on additional real problems
- Code (MIT License) is available by request, with the full release soon

ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.