GPEX, A Framework For Interpreting Artificial Neural Networks

NeurIPS 2023

Gaussian process, a good proxy model for interpreting artificial neural networks

Local explanation methods like LIME and SHAP

- are faithful to the original neural network only locally
- an adversary model can perturb a test instance to dramatically change the explanation provided by local methods [1]
- an adversary model can distinguish between and behave differently on the test instance itself and the perturbed versions, thereby fooling local explanation methods [2]

Gradient-based methods like DeepLift

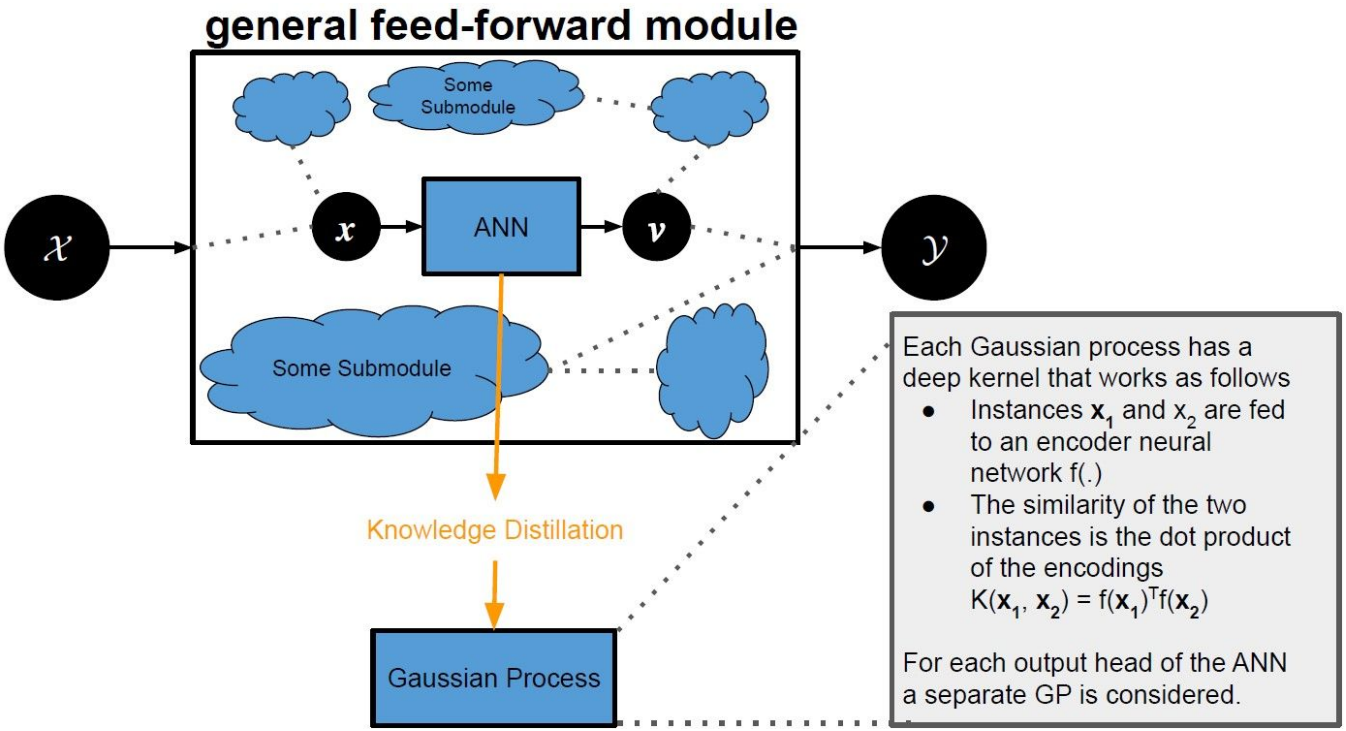
- have usually no theoretical backing
- different gradient-based methods may produce discordant explanations for a single test instance and a single neural network [3]

Gaussian process

- has the potential to become globally faithful to a neural network
- is a white-box model and is highly interpretable
- provides an uncertainty for every prediction that it makes

Finding a Gaussian process (GP) which is globally faithful to a neural network

- Previous approaches: make a neural network equivalent to a GP by, e.g., making each and every intermediate layer wide [4].
- Our approach: given an ANN (artificial neural network) find an equivalent and globally faithful GP via knowledge distillation.
  - Our formulation/implementation works for any neural network submodule of an arbitrary feed-forward module.
  - No theoretical assumption is made, however some of known assumptions may facilitate knowledge distillation.


general feed-forward module
Some Submodule
ANN
Some Submodule
Knowledge Distillation
Gaussian Process
Each Gaussian process has a deep kernel that works as follows
Instances $\mathbf{x}_1$ and x_2 are fed to an encoder neural network f(.)
The similarity of the two instances is the dot product of the encodings $K(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_1)^T f(\mathbf{x}_2)$
For each output head of the ANN a separate GP is considered.

Intuitive loss function to distill knowledge from neural network to Gaussian process

- Let's say the neural network submodule is replaced with Gaussian processes. The replaced Gaussian processes introduce new latent variables. Therefore, optimizing the training objective function of the whole pipeline requires inference.
- We assumed that the variational distribution is parametrized by the given neural network g(.). The derived variational lower bound relates the GP and the given neural network in an intuitive way, and is our objective function to distill knowledge from the neural network to GPs.

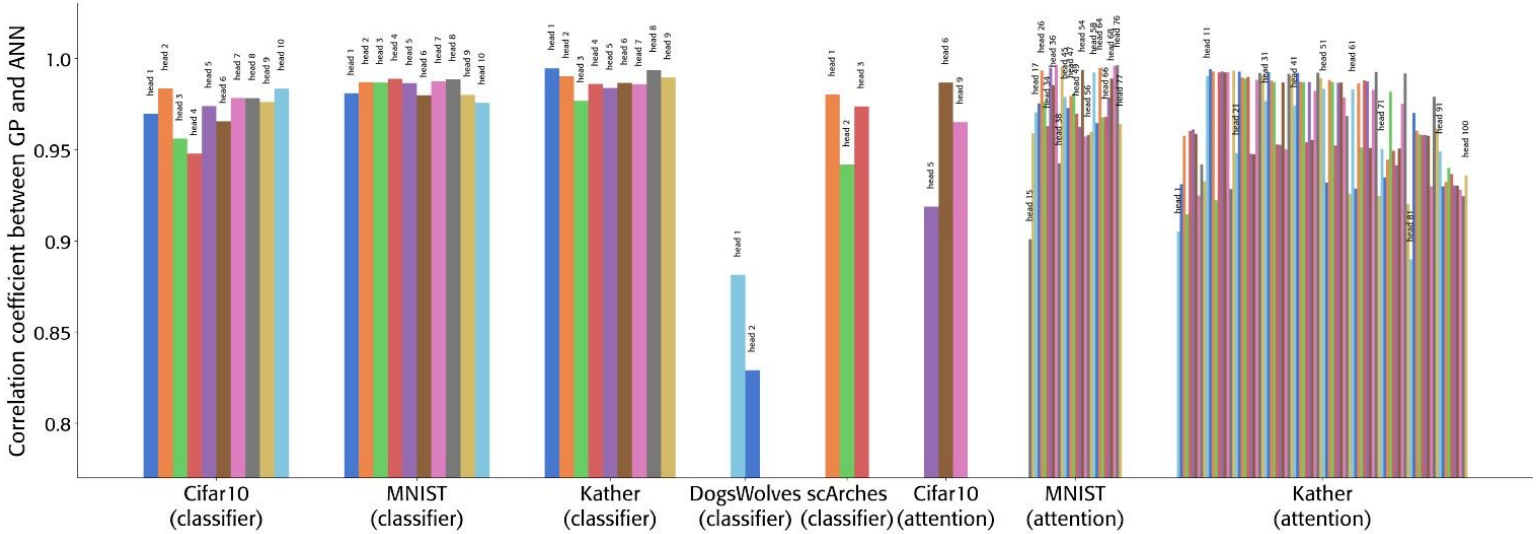
$$\mathcal{L}_{gp} = -\frac{1}{2}\mathbb{E}_{\sim q}\Big[\frac{\big(\mu_{gp}(\mathbf{x};.,.) - g(\mathbf{x})\big)^2}{cov_{gp}(\mathbf{x};.,.)}\Big] - \frac{1}{2}\mathbb{E}_{\sim q}\Big[\log(cov_{gp}(\mathbf{x};.,.))\Big] + \text{constants}$$

- In the above objective, when GP's uncertainty is high (resp. low), the denominator of the first term is big (resp. small) and the equivalence between the GP and neural network is less (resp. more) encouraged.

Matching the Gaussian processes to different neural networks

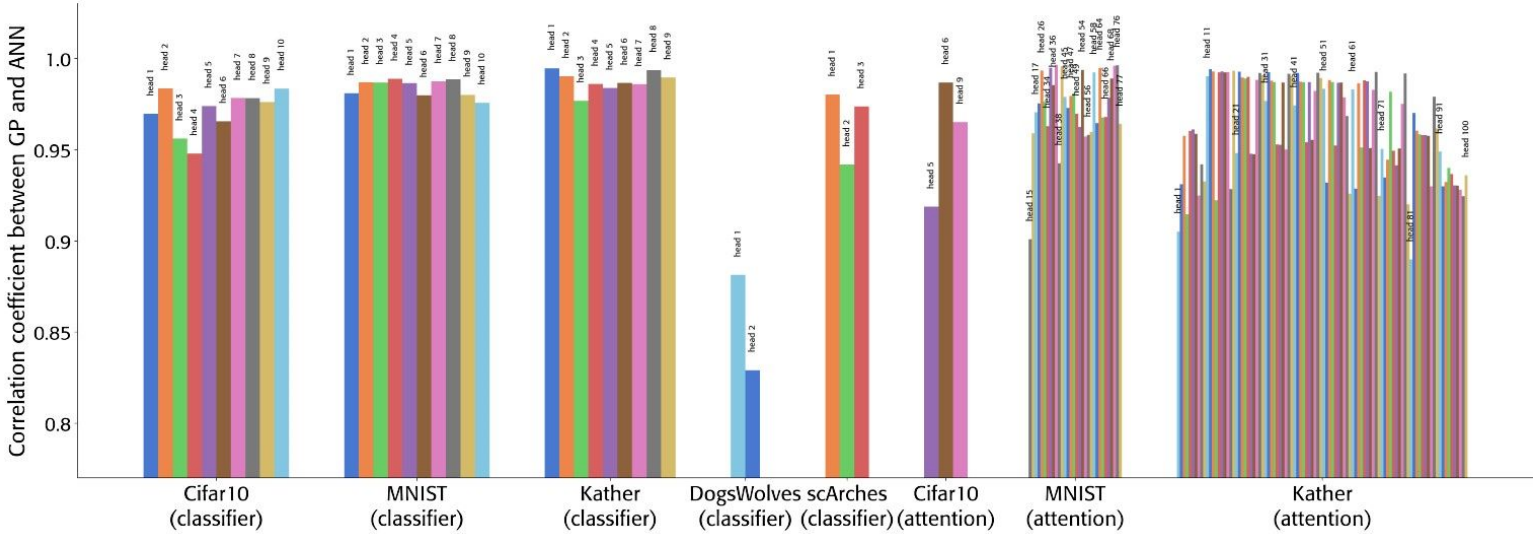
We applied our method to some neural networks

- ResNet classifiers, 1st group to the 4th group of bars in the below figure
- A feed-forward neural network that predicts the expression of some genes from cell embeddings, the 5th group of bars in the below figure.
- ResNet attention mechanism of attention-based classifier pipelines, 6th to 8th group of bars in the below figure.



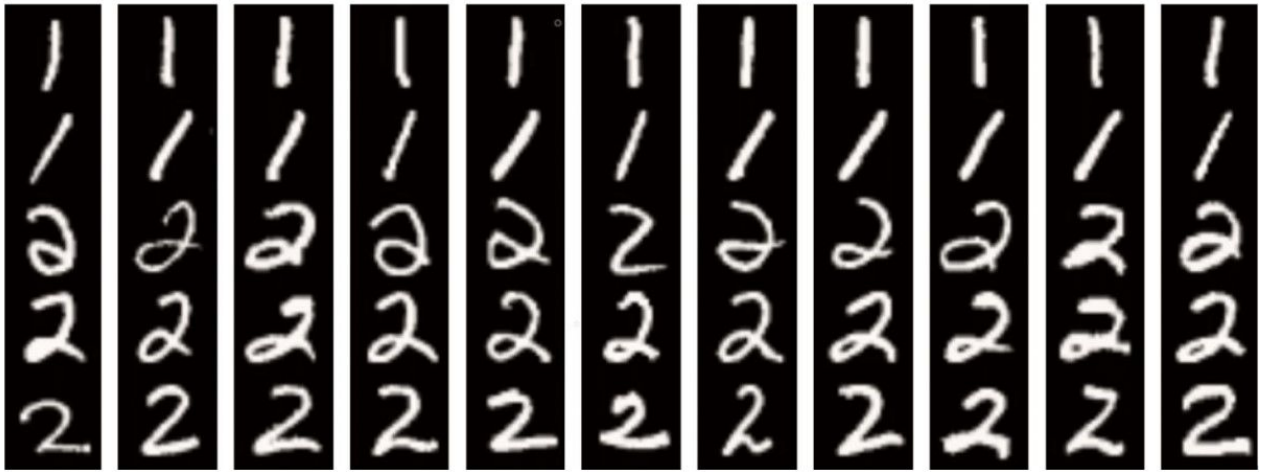
Matching the Gaussian processes to different neural networks

- In the above figure we see that the obtained GPs almost perfectly match the given neural networks.
- For the DogsWolves dataset (i.e. the 4th group of bars) the results are slightly lower, probably because the dataset has only 2000 instances.
- Our scalability techniques allow for including all training instances as inducing points for GPs, even if there are a million of instances. Our analysis on Cifar10 shows that having a lot of inducing points is crucial to get a perfect match between the GPs and neural networks.



Interpreting the decisions made by some neural network classifiers

For each classifier we looked at the similarity function of the obtained GPs. In the figures below the 1st column depicts a testing instance and columns 2-11 depict the 10 nearest neighbours to the test instance with counter-intuitive if not faulty focus/attention by the CNN as depicted by the matching heatmaps and despite good prediction. These findings are seen as opportunities to improve the learning process.

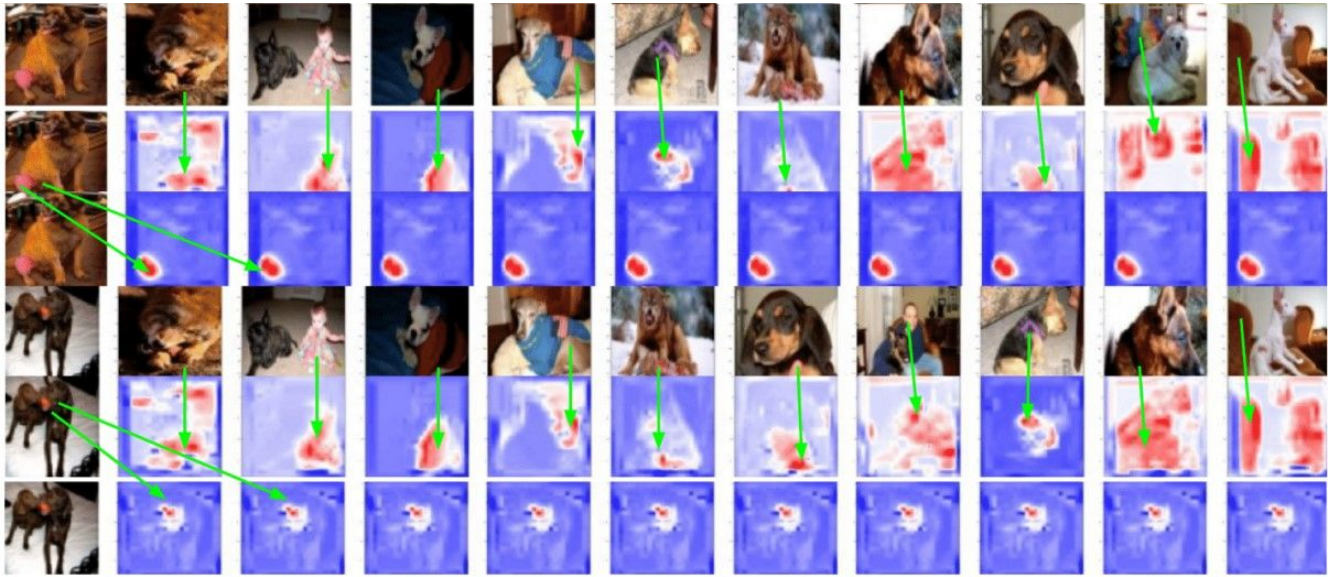


Interpreting the decisions made by some neural network classifiers

For each classifier we looked at the similarity function of the obtained GPs. In the figures below the 1st column depicts a testing instance and columns 2-11 depict the 10 nearest neighbours to the test instance with counter-intuitive if not faulty focus/attention by the CNN as depicted by the matching heatmaps and despite good prediction. These findings are seen as opportunities to improve the learning process.

Interpreting the decisions made by some neural network classifiers

For each classifier we looked at the similarity function of the obtained GPs. In the figures below the 1st column depicts a testing instance and columns 2-11 depict the 10 nearest neighbours to the test instance with counter-intuitive if not faulty focus/attention by the CNN as depicted by the matching heatmaps and despite good prediction. These findings are seen as opportunities to improve the learning process.



References

[1] Amirata Ghorbani et al. (2019). “Interpretation of neural networks is fragile.” In: Proceedings of the AAAI Conference on Artificial Intelligence, 33(01).

[2] Dylan Slack et al. (2020). “Fooling lime and shap: Adversarial attacks on post hoc explanation methods.”. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES ’20, page 180–186.

[3] Ashkan Khakzar et al. (2022). “Do explanations explain? model knows best.”. In: Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition, pages 10244–10253.

[4] Alexander Matthews et al. (2018). “Gaussian process behaviour in wide deep neural networks”. In: International Conference on Learning Representations.