Constrained Neural Networks for increased transparency

An-Phi Nguyen
PreDoc Researcher

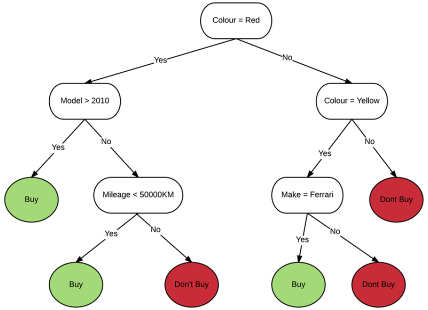
uye@zurich.ibm.com

@phineas_zu

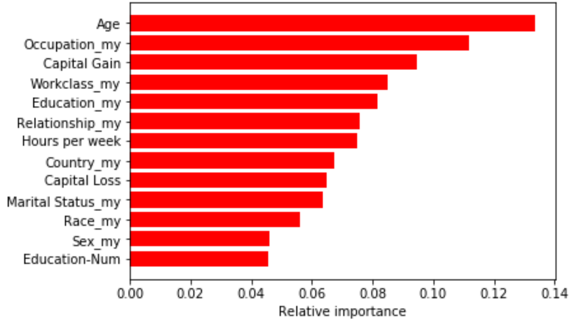
A quick glance on interpretability methods

Minimal taxonomy:

- "Time of design"



Ante-hoc



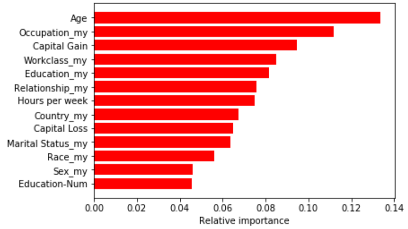
Post-hoc

A quick glance on Interpretability methods

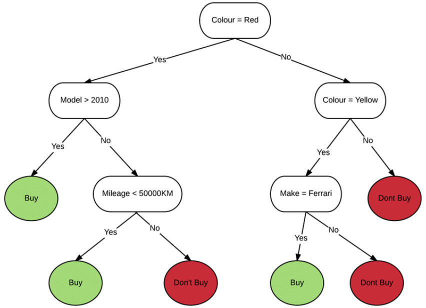
Minimal taxonomy:

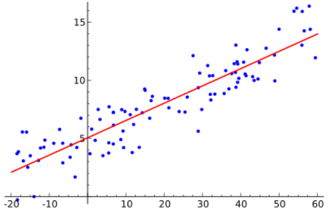
- Information visualization/amount

Feature attributions



Mechanistic/Algorithmic/Transparent/…

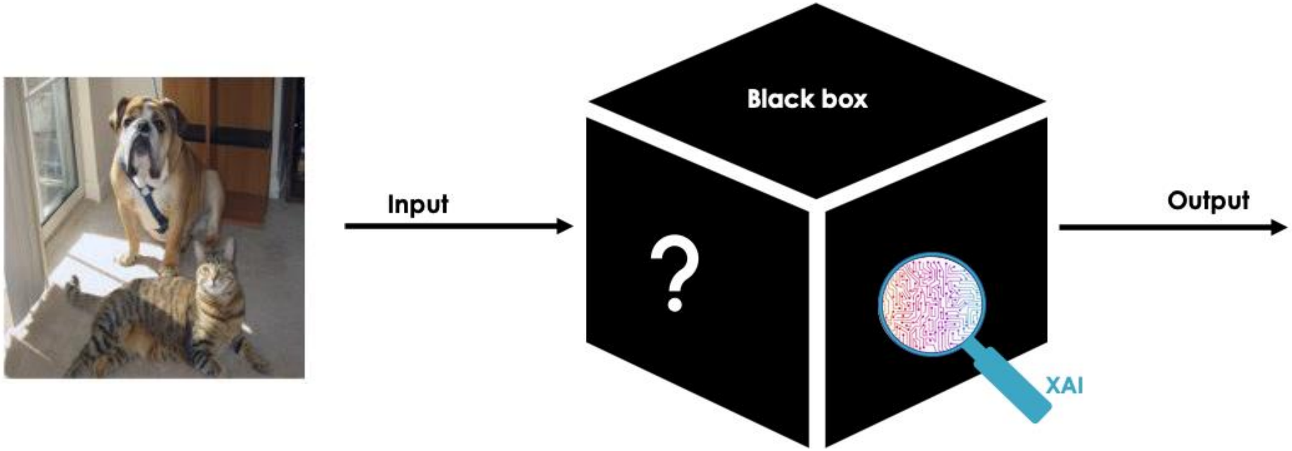




Example-based

Motivation

The recent trend consisted in training black-box models and later interpreted with post-hoc methods.



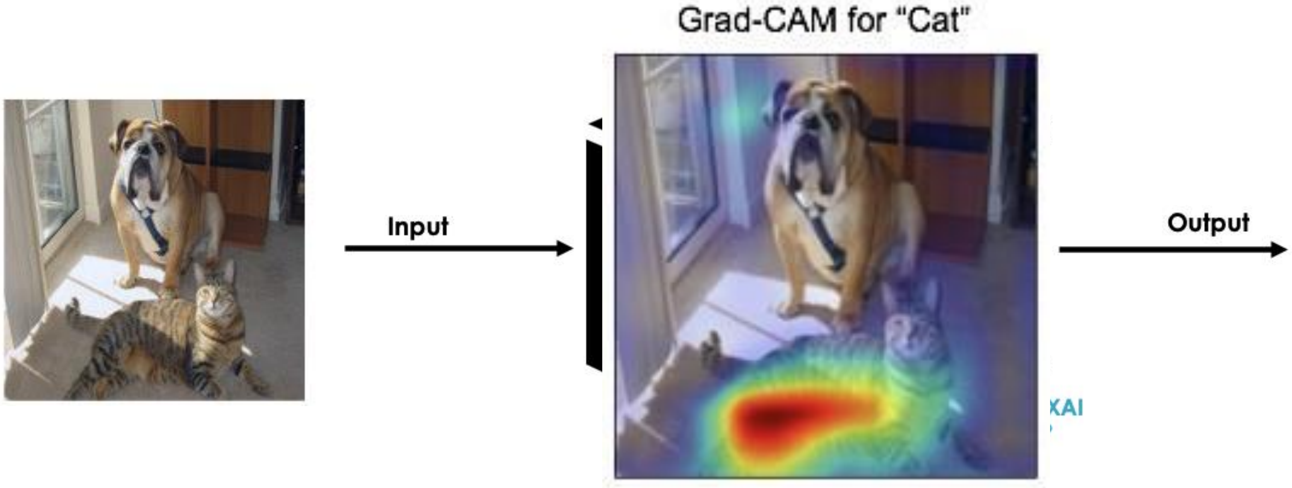
p_{dog}

p_{cat}

[1] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019)

Motivation

The recent trend consisted in training black-box models and later interpreted with post-hoc methods.



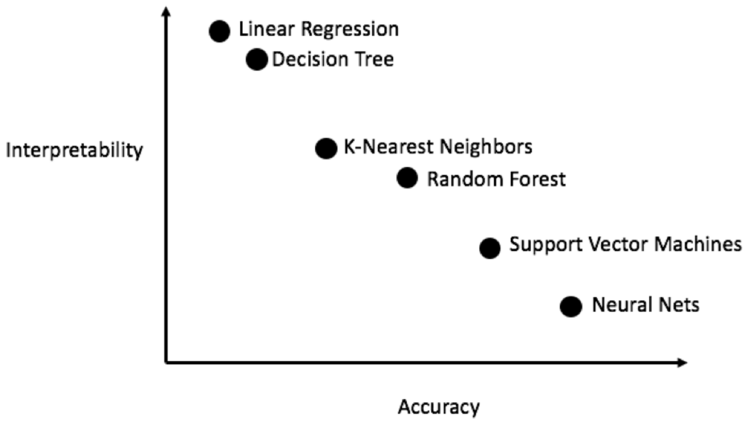
p_{dog}

p_{cat}

[1] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019)

Motivation

The recent trend consisted in training black-box models and later interpreted with post-hoc methods.

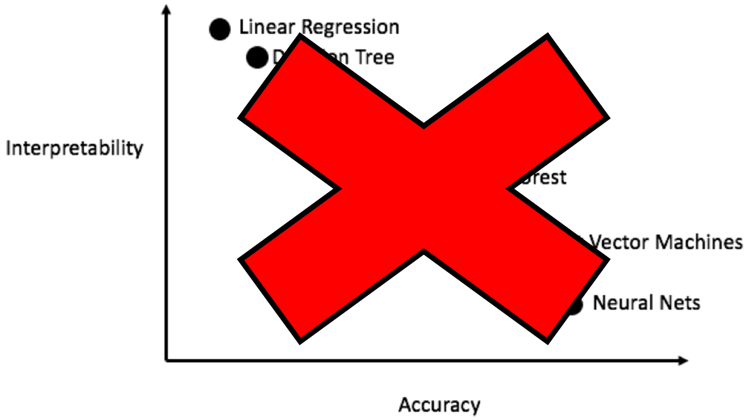


[1] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019)

Motivation

"Stop explaining black box machine learning models …

…use interpretable models instead" ~ Cynthia Rudin

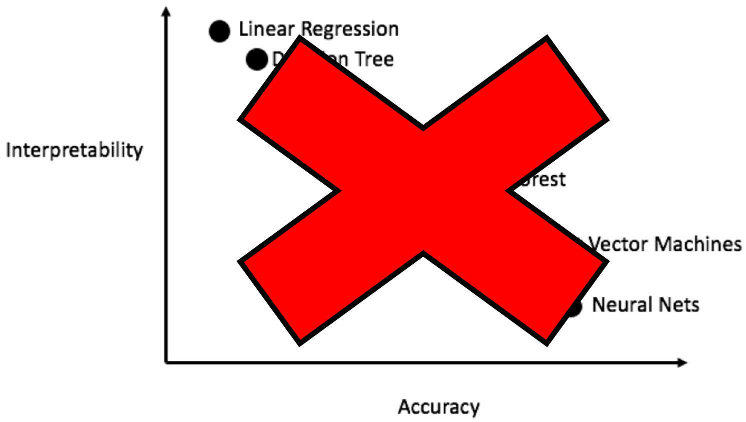


[1] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215 (2019)

Motivation

"For Tabular Data, additive models are enough"

~ Rich Caruana, yesterday.



[1] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)

Motivation

What about more complex domains?

- images, text,…

Motivation

What about more complex domains?

- Recent research proposes neural nets with some interpretability capabilities
  - Attention-based models [1]
  - Self-Explaining Neural Networks [2]
  - ProtoPNet [3]

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017)
[2] Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." *Advances in neural information processing systems* 31 (2018)
[3] Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *Advances in neural information processing systems* 32 (2019).

FLANs - Getting inspiration from Linear Models

Let's have a look at a *linear model*. They are usually considered easy to interpret. *Why*?

1. Separability
2. Predictability

$$y = b_0 + b_1x_1 + \ldots + b_Nx_N$$

FLANs - Getting inspiration from Linear Models

Let's have a look at a *linear model*. They are usually considered easy to interpret. *Why*?

1. Separability -> Modular interpretability
2. Predictability

$$y = b_0 + b_1x_1 + \dots + b_Nx_N$$

FLANs - Getting inspiration from Linear Models

Let's have a look at a *linear model*. They are usually considered easy to interpret. *Why*?

1. Separability -> Modular interpretability
2. Predictability -> We know the effect on the output

$$y = b_0 + b_1 x_1 + \ldots + b_N x_N$$

FLANs - Getting inspiration from Linear Models

Let's have a look at a *linear model*. They are usually considered easy to interpret. *Why*?

1. Separability -> Modular interpretability
2. Predictability -> Editable/Actionable

$$y = b_0 + b_1 x_1 + \dots + b_N x_N$$

FLANs - Getting inspiration from Linear Models

FLANs [1] extend this with two modifications:

1. Apply a non-linear function to *each feature*
2. Apply a non-linear function on the sum aggregation

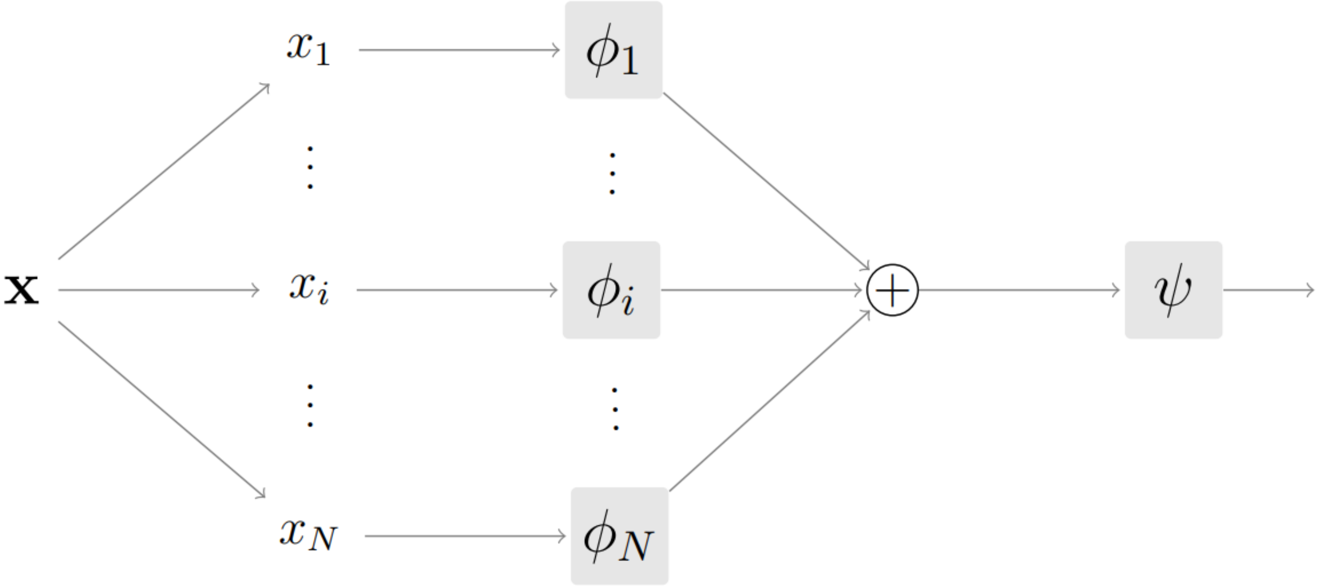
$$y = g(f_1(x_1) + \ldots + f_N(x_N))$$

[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs - Getting inspiration from Linear Models

FLANs [

1. App
2. App



[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs - Getting inspiration from Linear Models

FLANs extend this with two modifications:

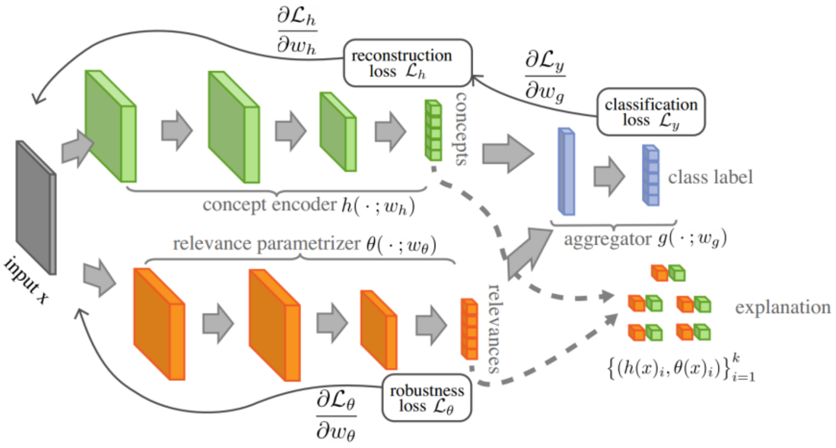
- both g and f are NNs trained via SGD or variants

$$y = g(f_1(x_1) + \ldots + f_N(x_N))$$

FLANs – Related Work - SENNs

Self-explaining networks [1]

- Linear aggregation as final layer
  - No separability
  - No predictability



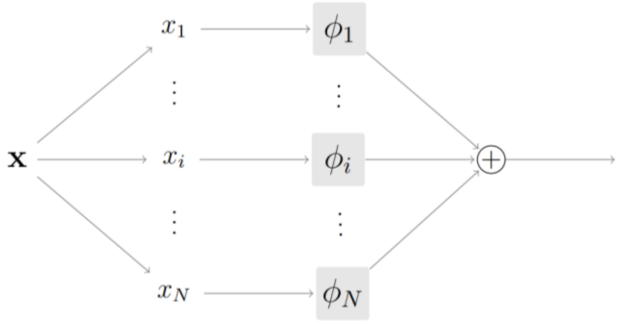
$$y = g_1(\mathbf{x})h_1(\mathbf{x}) + \ldots + g_N(\mathbf{x})h_N(\mathbf{x})$$

[1] Alvarez Melis, David, and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks." *Advances in neural information processing systems* 31 (2018)

FLANs – Related Work – EBMs and NAMs

EBMs [1] and Neural Additive Models [2]

- Additive aggregation
  - f is a decision tree in EBMs and a NN in NAMs
  - Exact separability
  - Exact predictability
    - But local expls not generalizable to global expls
  - But… No approximation power for complex data
    - Interactions have to be manually modeled



$$y = f_1(x_1) + \ldots + f_N(x_N) + \ldots$$

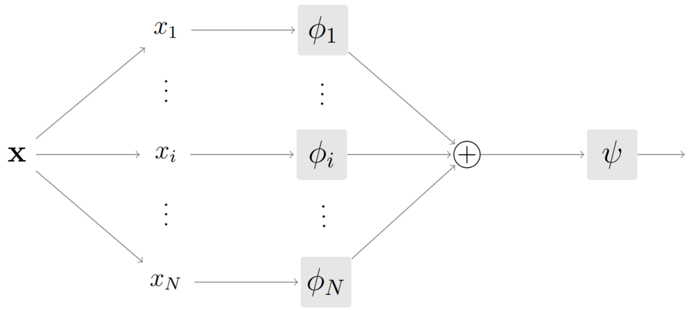
[1] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)

[2] Agarwal, Rishabh, et al. "Neural additive models: Interpretable machine learning with neural nets." *Advances in Neural Information Processing Systems* 34 (2021)

FLANs – Summary of the steps

The 3 steps of a FLAN model [1]:

1. **Map** *features* <u>*separately*</u> to a <u>*common*</u> latent space
2. **Sum** the feature representations
3. **Apply** another neural net for the final prediction

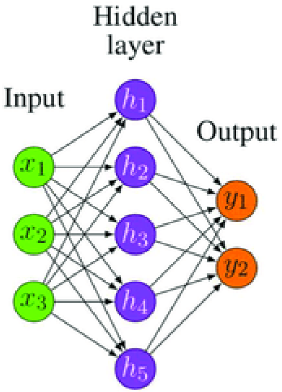


[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs – Universal approximators

Approximation capabilities given by the Kolmogorov-Arnold Representation Theorem [1]

- This same theorem is at the basis of the Approximation theorem for wide shallow neural nets

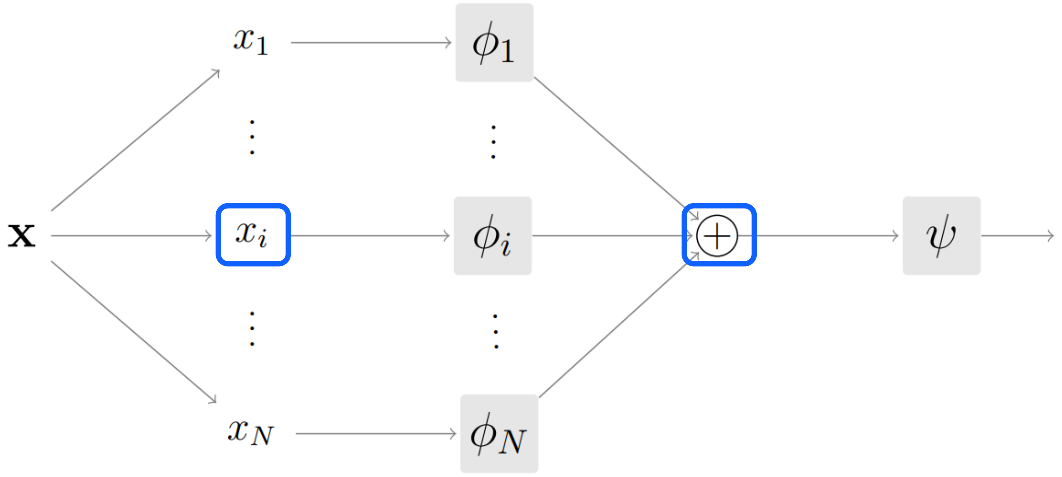


$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$

[1] Braun, J., Griebel, M. On a Constructive Proof of Kolmogorov's Superposition Theorem. *Constr Approx* **30,** 653 (2009)

FLANs – How do we interpret them?

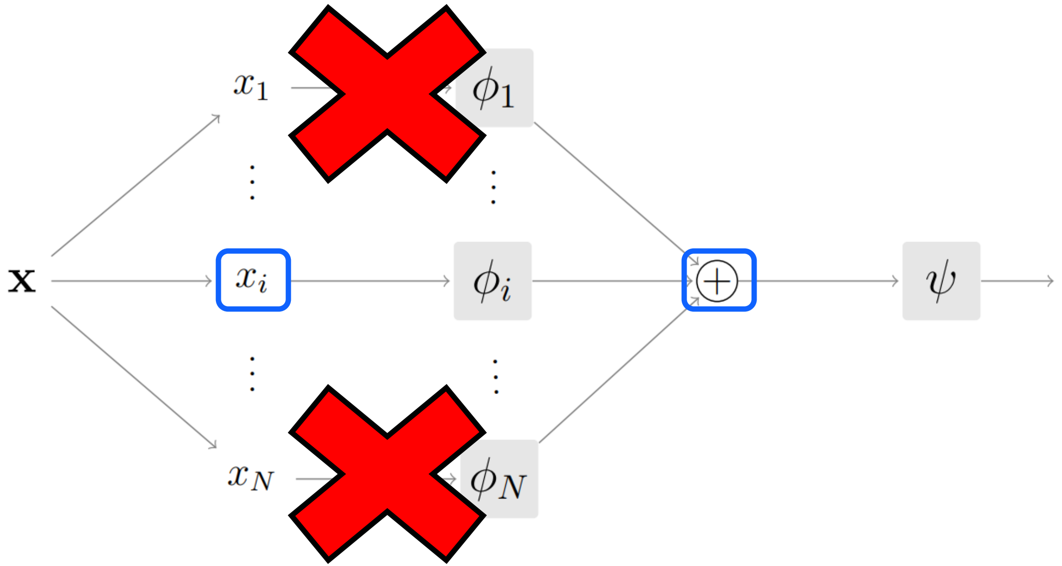
FLANs can be algorithmically interpreted similarly to additive models



[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs – How do we interpret them?

FLANs can be algorithmically interpreted similarly to additive models



[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs – How do we interpret them?

FLANs can be algorithmically interpreted similarly to additive models.. But approximately!

- We lose the exact predictability/separability
  - In exchange for higher accuracy/applicability on complex data

$$||\underbrace{\psi(\mathbf{z}_* + \mathbf{z}_i) - \psi(\mathbf{z}_*)}_{(\Delta)} - \psi(\mathbf{z}_i)||_{\mathcal{Y}} = ||\mathbf{J}_{\mathbf{z}_*}\mathbf{z}_i - \psi(\mathbf{z}_i) + o(||\mathbf{z}_*||_{\mathcal{Z}})||_{\mathcal{Y}}$$

[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs – How do we interpret them?

FLANs has some native way to compute importance… similar to attention scores.

- If a processed feature has almost-zero norm in latent space, it will not contribute to the final prediction
- -> norms are indicative of importance

[1] Nguyen, An-phi, and Maria Rodríguez Martínez. "It's FLAN time! Summing feature-wise latent representations for interpretability." *arXiv preprint arXiv:2106.10086* (2021)

FLANs – Remark

A feature can be anything user-defined

- a single feature
- hand-engineered features
- a group of features, e.g. patch

FLANs – Performance results – Tabular data

	COMPAS	adult	heart	mammo
Logistic Regression	0.905 (0.917) ± 0.006	0.892 (0.896) ± 0.003	0.873 (0.923) ± 0.032	**0.841** (0.874) ± 0.017
Decision Tree (small)	0.903 (0.915) ± 0.007	0.865 (0.871) ± 0.005	0.849 (0.882) ± 0.026	0.799 (0.818) ± 0.017
Decision Tree (unrestricted)	0.902 (0.915) ± 0.007	0.813 (0.821) ± 0.005	0.848 (0.882) ± 0.024	0.801 (0.826) ± 0.016
Random Forest	**0.915** (0.927) ± 0.007	0.869 (0.877) ± 0.004	0.945 (0.964) ± 0.014	0.822 (0.841) ± 0.016
EBM	0.911 (0.923) ± 0.008	**0.893** (0.896) ± 0.002	0.941 (0.959) ± 0.015	0.840 (0.869) ± 0.015
MLP	**0.915** (0.927) ± 0.006	0.874 (0.883) ± 0.005	0.937 (0.958) ± 0.023	0.831 (0.856) ± 0.014
SENN (Alvarez Melis and Jaakkola, 2018)	0.910 (0.922) ± 0.007	0.865 (0.873) ± 0.005	0.881 (0.925) ± 0.036	0.834 (0.860) ± 0.013
FLAN	0.914 (0.923) ± 0.004	0.880 (0.886) ± 0.004	**0.950** (0.973) ± 0.019	0.832 (0.867) ± 0.019

FLANs – Performance results – Images & Text

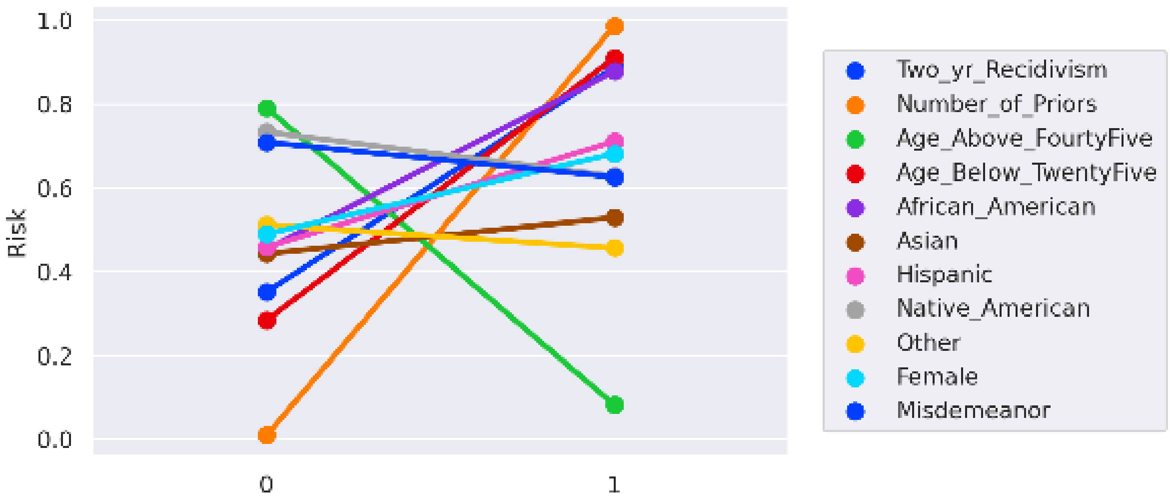
	MNIST	SVHN	CUB
ResNet	99.2	94.5*	84.5*
iCaps	99.2	92.0	-
ViT	-	88.9	90.4*
ProtoPNet	-	-	84.8*
SENN	99.1	-	-
SotA	**99.84**	**99.0***	**91.3***
FLAN	99.00 (99.05) $\pm$ 0.0007	93.37 (93.41) $\pm$ 0.0004	71.17 (71.53) $\pm$ 0.003

	AGNews	IMDb
CharCNN	90.49	-
LSTM	93.8	86.5
VDCNN	91.33	79.47
HAHNN	-	95.17
XLNet	**95.6***	**96.8***
FLAN	90.6 (90.9) $\pm$ 0.003	84.9 (85.1) $\pm$ 0.002

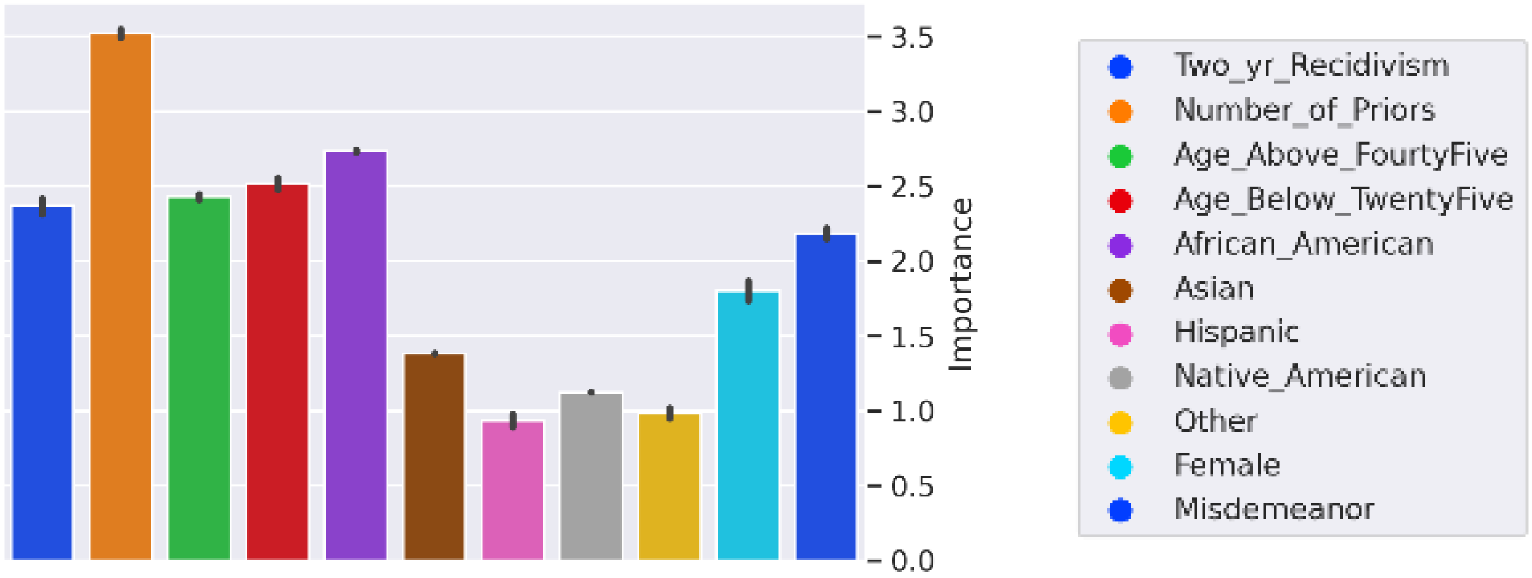
FLANs – Performance results – Take Aways

- SotA on tabular data
  - But our model is not needed on tabular data
  - Just a sanity check
- SotA on more complex datasets wrt interpretable models
- Lower accuracy wrt to unconstrained NNs
  - Do we need to model interactions?
  - Can we do a better architecture search?

FLANs – Some qualitative interpretability - COMPAS

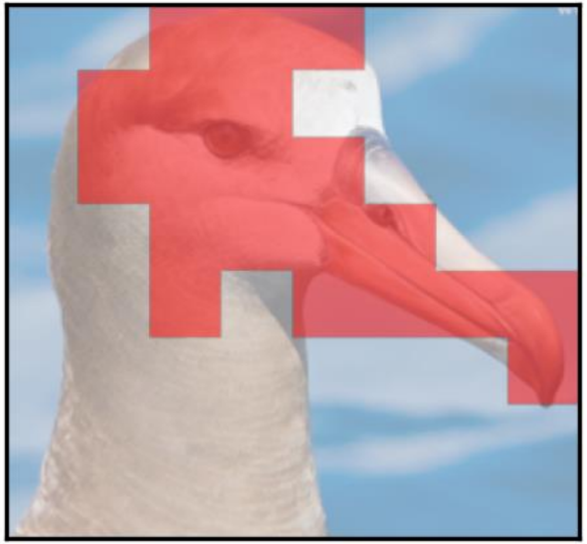


FLANs – Some qualitative interpretability - COMPAS



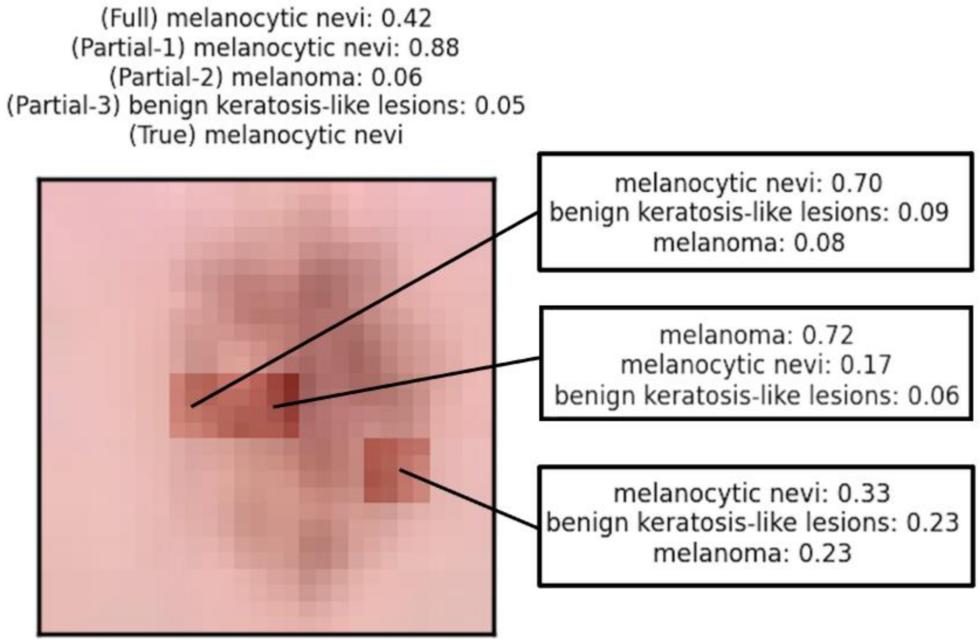
FLANs – Some qualitative interpretability - CUB

Black_footed_Albatross: 0.76
Laysan_Albatross: 0.12
Sooty_Albatross: 0.09



FLANs – Some qualitative interpretability - CUB

FLANs – Some qualitative interpretability – Skin Lesion



MonoNets – Monotonic constraints

- Monotonicity can be seen as an extension to linearity in some way
- The way to interpret it is a lil bit roundabout

[1] Nguyen, An-phi, and María Rodríguez Martínez. "Mononet: towards interpretable models by learning monotonic features." *arXiv preprint arXiv:1909.13611* (2019)

Summary

- For complex data/tasks, we need to trade-off accuracy vs. interpretability
- We can have an approximate *linear-like interpretability*
    - *The learned function is not linear itself!*

Future directions

- Find a better way to train them
  - Sum aggregation may be too restrictive in terms of learning
  - Should we reintroduce back some hand-engineered interaction?
  - Or is a better architecture search enough?
- Is the model really interpretable?
  - Linear interpretability seems appealing…
  - … but is it really useful or effective in complex scenarios?
  - User studies would be necessary

Thank you – Questions?

IBM