Can **Deep Learning** Be Interpreted with **Kernel Methods**?

Ben Edelman &
Preetum Nakkiran

Opening the black box of neural networks

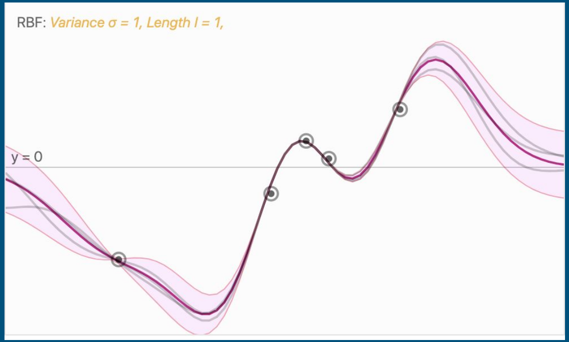
We've seen various post-hoc explanation methods (LIME, SHAP, etc.), but none that are faithful and robust.

Our view:

In order to generate accurate explanations, we need to leverage scientific/mathematical *understanding* of how deep learning works

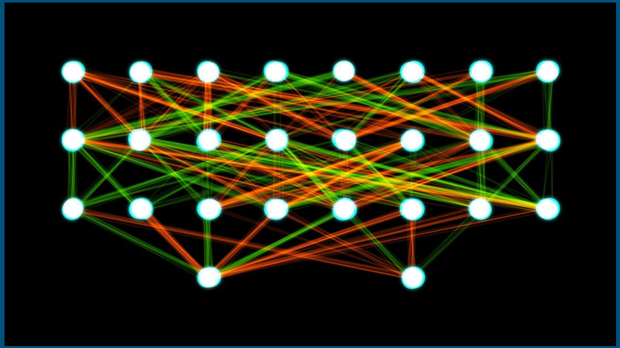
Kernel methods

- generalization guarantees
- closely tied to *linear regression*
- kernels yield interpretable similarity measures



Neural networks

- opaque
- no theoretical generalization guarantees



Equivalence: Random Fourier Features

Rahimi & Recht, 2007:

Training the final layer of a 2-layer network with cosine activations is *equivalent* (in large width limit) to running Gaussian kernel regression

- convergence holds empirically
- generalizes to any PSD shift-invariant kernel

Equivalence: Neural Tangent Kernel

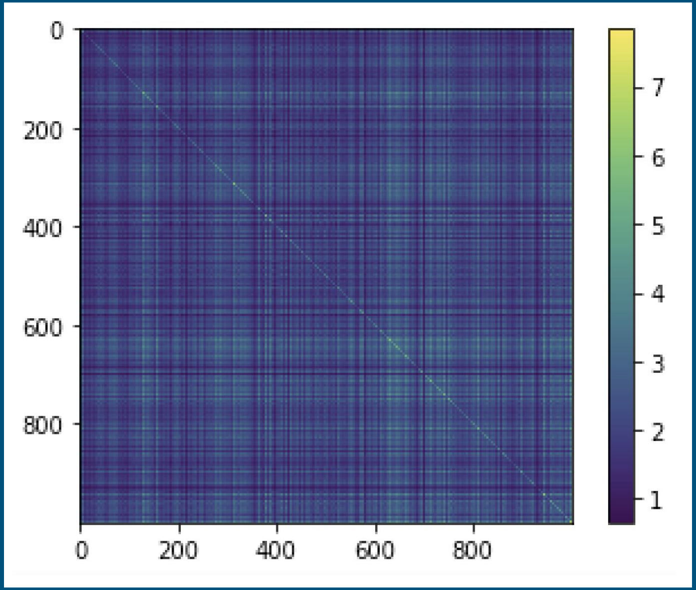
Jacot et al. 2018 & many follow-up papers:

Training a deep network (i.e. state-of-the-art conv. net) is *equivalent* (in the large width, small learning rate limit) to kernel regression with a certain corresponding "neural tangent kernel"

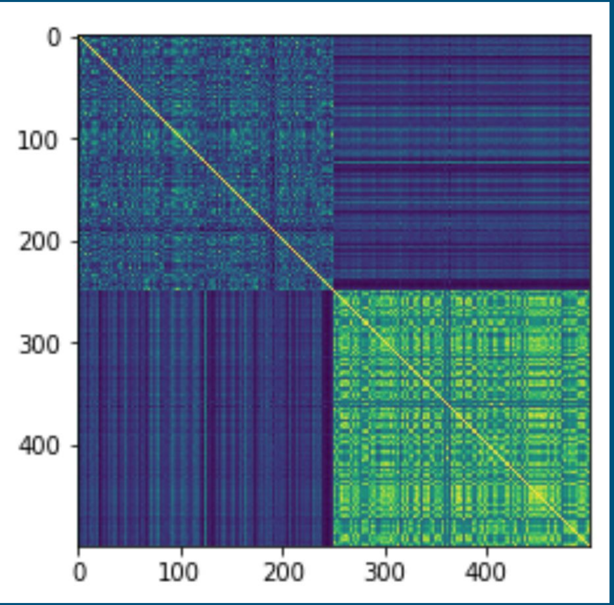
- but does the convergence hold empirically? (reasonable width)

Experiments

ENTK



Gaussian Kernel



Experiments

Q1: Why are RFFs (Gaussian Kernel) "well behaved" but not ENTK (for CNNs)?

Differences:

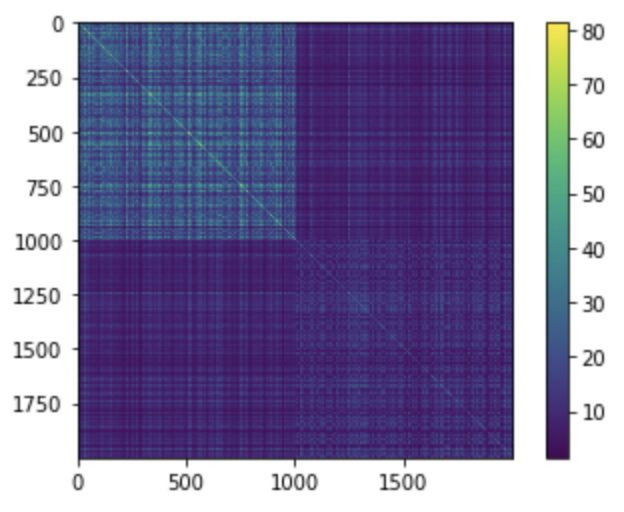
- Cosine vs. ReLU activation
- Architecture: deep CNN vs shallow fully-connected

Q2: Why is the Gaussian kernel interpretable?

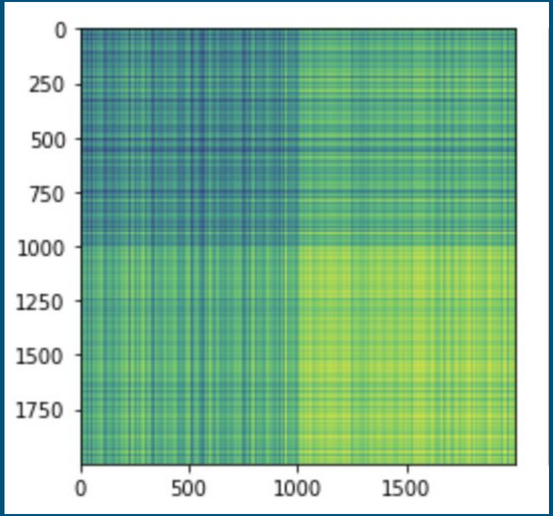
- Are there general properties that could apply to other kernels?

Q1: Relu vs Cosine activation

ReLU features



Cosine features



Q2: Why is Gaussian Kernel interpretable?

Experiment: Gaussian Kernel works on linearly-separable data (!)

Reason: Large-bandwidth gaussian kernel ~ "almost linear" embedding

$x \longrightarrow \sin(< w, x >) \sim \mathbf{<w, x>} - \frac{1}{2}(<w, x>)^2$

Conclusion

A question:

Can we find neural network architectures that are both

(a) *high performing* and
(b) correspond to "interpretable" kernels for reasonable widths?

Thank you!

Faithful and Customizable Explanations of Black Box Models

Lakkaraju, Kamar, Caruana, and Leskovec, 2019

Presented by: Christine Jou and Alexis Ross

Overview

I. Introduction
II. Framework
III. Experimental Evaluation
IV. Discussion

I. Introduction

A) Research Question
B) Contributions
C) Prior Work and Novelty

Research Question

How can we explain the **behavior** of black box classifiers **within specific feature subspaces**, while jointly **optimizing for fidelity, unambiguity, and interpretability**?

Contributions

- **Propose** *Model Understanding through Subspace Explanations* (MUSE), a new **model-agnostic framework** which explains black box models with decision sets that capture behavior in customizable feature subspaces.
- Create a **novel objective function** which jointly optimizes for *fidelity*, *unambiguity*, and *interpretability*.
- **Evaluate** the explanations learned from MUSE with experiments on **real-world datasets and user studies**.

Prior Work

- Visualizing and understanding specific models

- Explanations of model behavior:
  - ***Local* explanations** for individual predictions of a black box classifier (ex: LIME)
  - ***Global* explanations** for model behavior as a whole. Work of this sort has focused on approximating black box models with interpretable models such as decision sets/trees

Novelty

- **A new type of explanation:** *Differential* explanations, or global explanations within feature spaces of user interest, which allow users to explore how model logic varies within these subspaces
- Ability to **incorporate user input** in explanation generation

II. Framework

Model Understanding through Subspace Explanations (MUSE)

A) Workflow
B) Representation
C) Quantifying Fidelity, Unambiguity, and Interpretability
D) Optimization

Workflow

1) Design representation
2) Quantify notions
3) Formulate optimization problem
4) Solve optimizing efficiently
5) Customize explanations based on user preferences

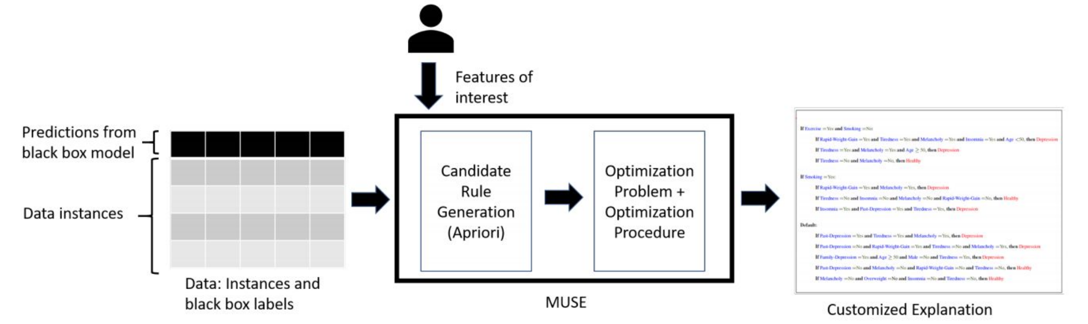
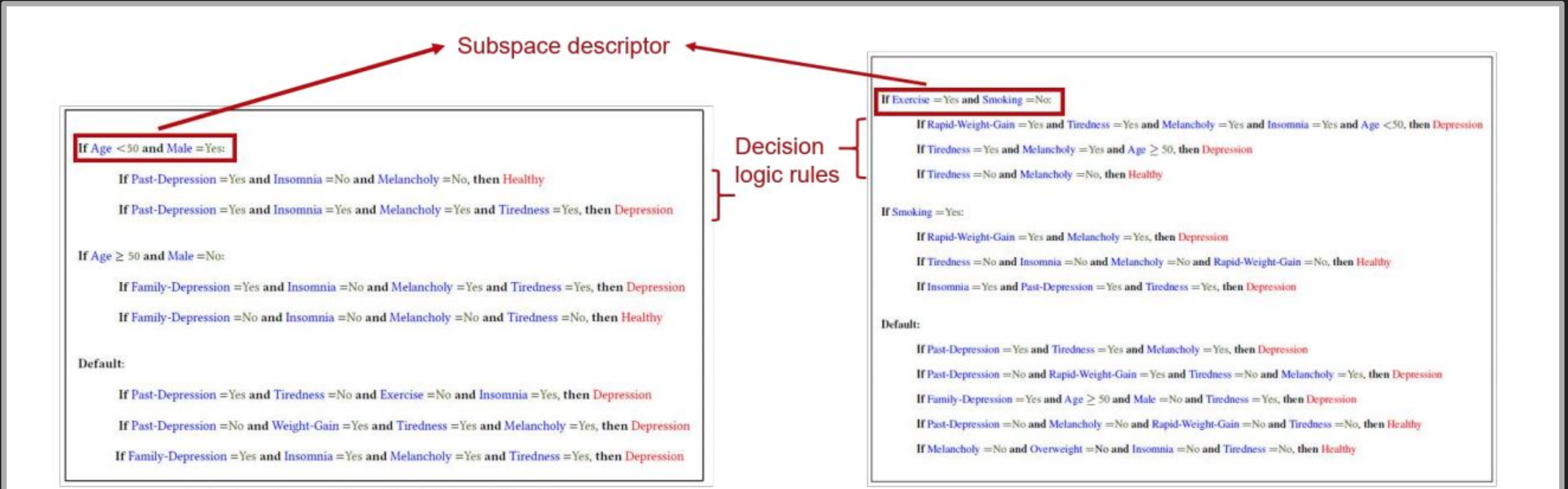


Figure 2: Algorithmic flow of MUSE approach: MUSE takes data, black box model predictions and user's features of interest. It outputs customized explanations.

Example of Generated Explanations



Explanation of the Black Box Model
(No user input)

Explanation of the Black Box Model
w.r.t. **Exercise & Smoking**

Representation: Two Level Decision Sets

- **Most important criterion for choosing representation list:** should be understandable to decision makers who are not experts in machine learning
- Two Level Decision Set
  - Basic building block of if-then rules that is unordered
  - Can be regarded as a set of multiple decision sets
- Definitions:
  - **Subspace descriptors:** conditions in the outer if-then rules
  - **Decision logic rules:** inner if-then rules

Important for **incorporating user input** and **describing subspaces** that are areas of interest

What is a Two-Level Decision Set?

Two Level Decision Set R is a set of rules $\{(q_1, s_1, c_1), (q_2, s_2, c_2), \ldots(q_M, s_M, c_M)\}$ where q_i and s_i are conjunctions of predicates of the form (feature, operator, value) and ci is a class label (*i.e. age > 50*)

- q_i corresponds to the subspace descriptor
- (s_i, c_i) together represent the inner if-then rules with s_i denoting the condition and c_i denoting the class label

A label is assigned to an instance x as follows:

- If x satisfies exactly one of the rules, then its label is the corresponding class label c_i
- If x satisfies none of the rules in R, then its label is assigned using the default function
- If x satisfies more than one rule in R then its label is assigned using a tie-breaking function

Quantifying Fidelity, Unambiguity, and Interpretability

- **Fidelity:** Quantifies disagreement between the labels assigned by the explanation and the labels assigned by the black box model
    - ***Disagreement(R):*** number of instances for which the label assigned by the black box model B does not match the label c assigned by the explanation R
- **Unambiguity:** Explanation should provide unique deterministics rationales for describing how the black box model behaves in various parts of the feature space
    - ***Ruleoverlap(R):*** captures the number of additional rationales provided by the explanation R for each instance in the data (higher values → higher ambiguity)
    - ***Cover(R):*** captures the number of instances in the data that satisfy some rule in R
    - Goal: minimize **ruloverlap(R)** and **maximize cover(R)**

Quantifying Fidelity, Unambiguity, and Interpretability (cont.)

- **Interpretability:** Quantifies how easy it is to understand and reason about explanation (often depends on complexity)
  - ***Size(R):*** number of rules (triples of the form (q,s,c)) in the two level decision set R
  - ***Maxwidth(R):*** maximum width computed over all the elements in R where each element is either a condition of some decision logic rule s or a subspace descriptor q, where width(s) is the number of predicates in the condition x
  - ***Numpreds(R):*** the number of predicates in R including those appearing in both the decision logic rules and subspace descriptors
  - ***Numdsets(R):*** the number of unique subspace descriptions (outer if-then clauses) in R

Formalization of Metrics

- Subspace descriptors and decision logic rules have ***different*** semantic meanings!
    - Each subspaces descriptor characterizes a specific region of the feature space
    - Corresponding inner if-then rules specify the decision logic of the black box model within that region
- We want to minimize the overlap between the features that appear in the subspace descriptors and those that appear in the decision logic rules

Formalization of Metrics

Table 1: Metrics used in the Optimization Problem

Fidelity	$disagreement(\mathcal{R}) = \sum_{i=1}^{M} \lvert\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}, \boldsymbol{x} \text{ satisfies } q_i \wedge s_i, \mathcal{B}(\boldsymbol{x}) \neq c_i\}\rvert$
Unambiguity	$ruleoverlap(\mathcal{R}) = \sum_{i=1}^{M} \sum_{j=1, i \neq j}^{M} overlap(q_i \wedge s_i, q_j \wedge s_j)$ $cover(\mathcal{R}) = \lvert\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}, \boldsymbol{x} \text{ satisfies } q_i \wedge s_i \text{ where } i \in \{1 \cdots M\}\}\rvert$
Interpretability	$size(\mathcal{R})$: number of rules (triples of the form (q, s, c)) in $\mathcal{R}$ $maxwidth(\mathcal{R}) = \max_{e \in \bigcup_{i=1}^{M} (q_i \cup s_i)} width(e)$ $numpreds(\mathcal{R}) = \sum_{i=1}^{M} width(s_i) + width(q_i)$ $numdsets(\mathcal{R}) = \lvert dset(\mathcal{R})\rvert \text{ where } dset(\mathcal{R}) = \bigcup_{i=1}^{M} q_i$ $featureoverlap(\mathcal{R}) = \sum_{q \in dset(\mathcal{R})} \sum_{i=1}^{M} featureoverlap(q, s_i)$

Optimization

- **Objective Function:** non-normal, non-negative, submodular, and the constraints of the optimization problem are matroids

$$
\begin{aligned}
f_1(\mathcal{R}) &= \mathcal{P}_{max} - numpreds(\mathcal{R}), \text{ where } \mathcal{P}_{max} = 2 * \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}| \\
f_2(\mathcal{R}) &= \mathcal{O}_{max} - featureoverlap(\mathcal{R}), \text{ where } \mathcal{O}_{max} = \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}| \\
f_3(\mathcal{R}) &= \mathcal{O}'_{max} - ruleoverlap(\mathcal{R}), \text{ where } \mathcal{O}'_{max} = N \times (|\mathcal{ND}| * |\mathcal{DL}|)^2 \\
f_4(\mathcal{R}) &= cover(\mathcal{R}) \\
f_5(\mathcal{R}) &= \mathcal{F}_{max} - disagreement(\mathcal{R}), \text{ where } \mathcal{F}_{max} = N \times |\mathcal{ND}| * |\mathcal{DL}|
\end{aligned}
$$

$$
{}_{\mathcal{R} \subseteq \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}} \sum_{i=1}^{5} \lambda_i f_i(\mathcal{R})
$$

$$
\text{s.t. } size(\mathcal{R}) \leq \epsilon_1,\ maxwidth(\mathcal{R}) \leq \epsilon_2,\ numdsets(\mathcal{R}) \leq \epsilon_3
$$

ND: candidate set of predicates for subspace descriptors

DL: candidate set of predicates for decision logic rules

W_{max}: maximum width of any rule in either candidate sets

Optimization (cont.)

- **Optimization Procedure**
  - NP-hard
  - *Approximate local search:* provides the best known theoretical guarantees
- **Incorporating User Input**
  - User inputs a set of features that are of interest → workflow restricts the candidate set of predicates ND from which subspace descriptors are chosen
  - Ensures that the subspaces in the resulting explanations are characterized by the features of interest
  - **Featureoverlap(R)** and **f_2(R)** of objective function ensure that features that appear in subspace descriptors do not appear in the decision logic rules
- **Parameter tuning:**
  - Use validation set (5% of total data)
  - Initialize ƛ values to 100 and carry out coordinate descent style
  - Use apriori with 0.1 support threshold to generate candidates for conjunctions of predicates

Optimization (cont.)

Solution set initially empty

Delete and/or exchange operations until no other element remaining to be deleted or exchanged

Repeat k+1 times and return solution set with maximum value

Algorithm 1: Optimization Procedure [6]

```
Input: Objective f, domain ND × DL × C, parameter δ, number of constraints k
V_1 = ND × DL × C
for i ∈ {1, 2 ··· k + 1} do                ▷ Approximation local search procedure
    X = V_i; n = |X|; S_i = ∅
    Let v be the element with the maximum value for f and set S_i = v
    while there exists a delete/update operation which increases the value of S_i by a factor of at least (1 + δ/n^4) do
        Delete Operation: If e ∈ S_i such that f(S_i\{e}) ≥ [(1 + δ/n^4)] f(S_i), then S_i = S_i\e

        Exchange Operation If d ∈ X\S_i and e_j ∈ S_i (for 1 ≤ j ≤ k) such that
        (S_i\e_j) ∪ {d} (for 1 ≤ j ≤ k) satisfies all the k constraints and
        f(S_i\{e_1, e_2 ··· e_k} ∪ {d}) ≥ [(1 + δ/n^4)] f(S_i), then S_i = S_i\{e_1, e_2, ··· e_k} ∪ {d}
    end while
    V_{i+1} = V_i\S_i
end for
return the solution corresponding to max{f(S_1), f(S_2), ··· f(S_{k+1})}
```

Optimization (cont.)

Algorithm 1: Optimization Procedure [6]

Input: Objective f, domain $\mathcal{ND} \times \mathcal{DL} \times C$, parameter δ, number of constraints k

$V_1 = \mathcal{ND} \times \mathcal{DL} \times C$

for $i \in \{1, 2 \cdots k+1\}$ **do** ▷ Approximation local search procedure

  $X = V_i$; $n = |X|$; $S_i = \emptyset$

  Let v be the element with the maximum value for f and set $S_i = v$

  while there exists a delete/update operation which increases the value of S_i by a factor of at least $(1 + \frac{\delta}{n^4})$ **do**

    Delete Operation: If $e \in S_i$ such that $f(S_i \backslash \{e\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash e$

    Exchange Operation If $d \in X \backslash S_i$ and $e_j \in S_i$ (for $1 \leq j \leq k$) such that

    $(S_i \backslash e_j) \cup \{d\}$ (for $1 \leq j \leq k$) satisfies all the k constraints and

    $f(S_i \backslash \{e_1, e_2 \cdots e_k\} \cup \{d\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash \{e_1, e_2, \cdots e_k\} \cup \{d\}$

  end while

  $V_{i+1} = V_i \backslash S_i$

end for

return the solution corresponding to $\max\{f(S_1), f(S_2), \cdots f(S_{k+1})\}$

III. Experimental Evaluation

A) Experimentation with Real World Data
B) Evaluating Human Understanding of Explanations with User Studies

Experimentation with Real World Data

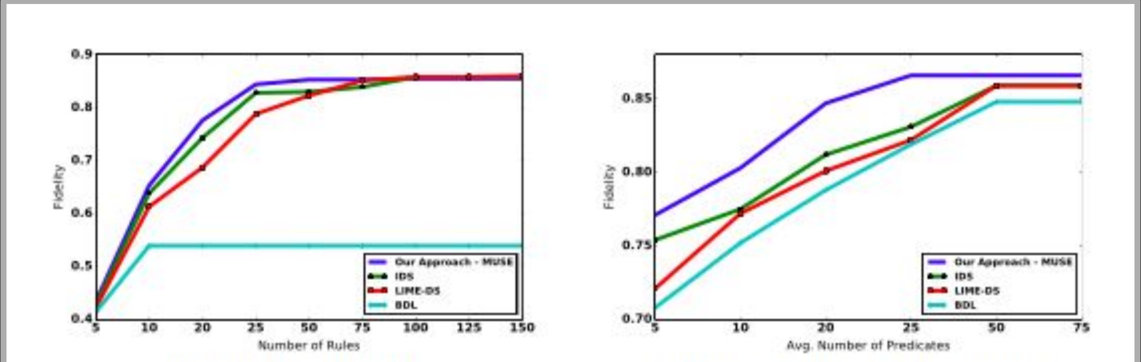
- Compare the quality of explanations generated by MUSE with quality of explanations generated by other state-of-the-art baselines
  - Fidelity vs. interpretability trade-offs
  - Unambiguity of explanations

Experimentation with Real World Data: Set-Up

- **Datasets:**
  1) Dataset of **bail outcomes**
  2) Dataset of high school **student performance** records
  3) **Depression diagnosis** dataset
- **Baselines:**
  - Decision set version of LIME (LIME-DS)
  - Interpretable Decision Sets (IDS)
  - Bayesian Decision Lists (BDL)
- **Model:** Deep neural network with 5 layers
- Treat model predictions as ground truth labels and approximate them

Experimentation with Real World Data: Set-Up

- *Fidelity* vs. Interpretability [# of rules (*size*), avg. # of predicates (*numpreds*)]
  - Results for the depression data set:



(a) Number of Rules
(b) Avg. Number of Predicates

  - MUSE explanations provide better trade-offs of fidelity vs. interpretability compared to other baselines

Experimentation with Real World Data: Set-Up

- Unambiguity of Explanations
  - Evaluate using *ruleoverlap* and *cover*
  - Results: MUSE-generated explanations result in low values of *ruleoverlap* (between 1% and 2%) and high values for *cover* (95% to 98%)

User Studies to EValuate Human Understanding of Explanations

- **Model:** 5 layer deep neural network
- **Data:** Depression diagnose dataset
- Evaluate the understanding MUSE-explanations offer users about black box models

User Study 1

- Question: What kind of understanding do different explanations provide users of how models behave in different parts of feature space?
- 33 participants
- Each participant randomly presented with explanations generated by:
    - MUSE, IDS, BDL
- Participants asked 5 questions about model behavior in different subspaces of feature space
    - Example: *Consider a patient who is female and aged 65 years. Based on the approximation shown above, can you be absolutely sure that this patient is Healthy? If not, what other conditions need to hold for this patient to be labeled as Healthy?*
- Computed **accuracy** of answers and **time taken to answer** each question

User Study 1: Results

- *Results:*
    - Users more accurate with explanations produced by MUSE (than by IDS or BDL)
    - Users about 1.5 (IDS) and 2.3 (BDL) times faster when using MUSE-generated explanations

Approach	Human Accuracy	Avg. Time (in secs.)
MUSE (No customization)	94.5%	160.1
IDS	89.2%	231.1
BDL	83.7%	368.5

User Study 2

- Question: What is the benefit obtained when the explanation presented to the user is customized with regard to the question the user is trying to answer?
- Ask the same 5 questions as before, but show user an explanation where the features being asked about appear in the subspace descriptors
  - Example: *Consider a patient who is female and aged 65 years. Based on the approximation shown above, can you be absolutely sure that this patient is Healthy? If not, what other conditions need to hold for this patient to be labeled as Healthy?*
    → Exercise and smoking would appear in the subspace descriptors, simulating the effect of the user inputting these features to customize the explanation
- 11 participants

User Study 2: Results

- Time taken to answer questions halved compared to setting where MUSE explanations were not customized
- Answers also slightly more accurate

Approach	Human Accuracy	Avg. Time (in secs.)
MUSE (No customization)	94.5%	160.1
MUSE (Customization)	98.3%	78.3

User Study 3

- <u>Question:</u> How do MUSE explanations compare with LIME explanations?
- Online survey where participants were shown MUSE explanations and LIME explanations and asked which they would **prefer to use** to answer questions of the previously mentioned form
- 12 participants

- <u>Results:</u> “Unanimous preference for MUSE explanations”

IV. Discussion

A) Conclusions and Discussion
B) Themes from comments

Conclusions

- Explanations generated using the MUSE framework are more “customizable, compact, easy-to-understand, and accurate” than explanations generated with other state of the art methods
- Future research directions:
  - Combine framework with efforts to extract interpretable features from images
  - Notions of fidelity, unambiguity, and interpretability could be further developed to account for certain features being more interpretable than others

Discussion: Our Notes

- Limited number of participants/questions in user studies
- Comparisons with LIME are limited
- Interactivity: is one explanation or multiple explanations better?
- No experiments investigating the contributions of each term of the objective function
- No experiments testing for whether MUSE explanations give *global* understanding of model. Do explanations help users:
    - select the best (unbiased) classifier?
    - improve a classifier by removing features that do not generalize?
    - trust a classifier?
    - have insights into a classifier?

Themes from Your Comments

Novelty: intuitive representation but less unique?

Metrics in user study: are there other metrics than speed and accuracy worth evaluating?

Higher order decision sets: would the results still translate to the same findings on different orders?

Diverse datasets: size of data? data with less priors?

Interactivity: one explanation or multiple explanations?

THANK YOU!

Questions?