GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction

Thai Le, Suhang Wang, Dongwon Lee

The Pennsylvania State University

August 23-27, 2020

KDD2020

OUTLINE

1. ✓ Introduction
2. Motivation & Challenges
3. Contrastive Explanation
4. GRACE Algorithm
5. Experiments
6. Conclusion

Motivation - Previous Methods

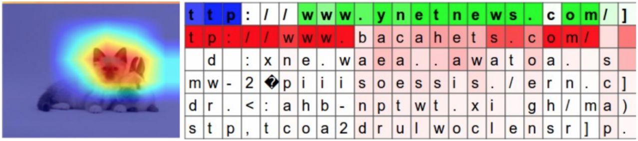

Figure: Example of Highlighting Spans of Words/Phrases. Source: Google Image

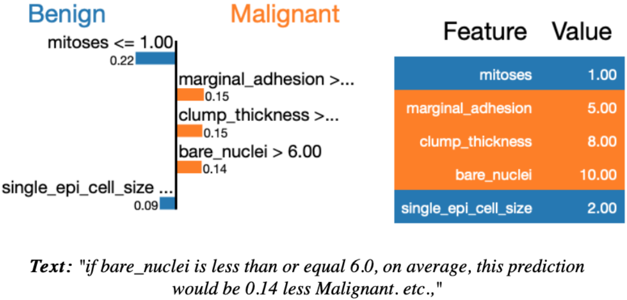

Figure: Example of Lime [Ribeiro et al., 2016] Method on Tabular Data Instance

Motivation - Challenges

1. **Tabular data** is prominent in many important fields.
2. Explanation for tabular data: high-dimensional inter-correlated features
   - Which key features to select?
   - If top K, what if features are highly correlated? (E.g. "Frequency of !" and "Frequency of !!!")
3. How to present the explanation to the end-users?
   - Highlighting a batch of an image, a span of words in a sentence?
4. **End-users** might have different interests than researchers, developers
5. **GOAL: Develop an algorithm to explain neural network models' predictions on tabular datasets to end-users**

OUTLINE

1. ✓ Introduction
2. ✓ Motivation & Challenges
3. Contrastive Explanation
4. GRACE Algorithm
5. Experiments
6. Conclusion

Contrastive Sample - Definition

Feature	freq_now	freq_credit	freq_!!!	freq_!	class
$\boldsymbol{x}_1$	0.1	0.0	0.0	0.0	Ham
$\tilde{\mathbf{x}}_1$	0.1	0.0	**0.3**	**0.453**	**Spam**

Feature	freq_you	freq_direct	avg_longest_capital	class
$\boldsymbol{x}_2$	0.68	0.34	158.0	Spam
$\tilde{\mathbf{x}}_2$	0.68	0.34	**1.0**	**Ham**

Table: Examples of original samples $\boldsymbol{x}_i$ and contrastive samples $\tilde{\boldsymbol{x}}_i$ on *spam* dataset.)

Contrastive Explanation - Definition

1. End-users are interested in explanation: "Why X rather than Y?"

Feature	freq_you	freq_direct	avg_longest_capital	class
$\boldsymbol{x}$	0.68	0.34	158.0	Spam
$\tilde{\mathbf{x}}$	0.68	0.34	**1.0**	**Ham**

Examples

Explanation *"Had the message had no words written in all capital letters, it would have been classified as **ham rather than spam."***

Explanation by Intervention

Problem Statement

Given $\boldsymbol{x}$ and neural network model $f(\cdot)$, our goal is to generate new contrastive sample $\tilde{\mathrm{x}}$ to provide concise and informative explanation for the prediction $f(\boldsymbol{x})$.

OUTLINE

1. ✓ Introduction
2. ✓ Motivation & Challenges
3. ✓ Contrastive Explanation
4. GRACE Algorithm
5. Experiments
6. Conclusion

Objective Formulation

Feature	freq_now	freq_credit	freq_!!!	freq_!	class
$\boldsymbol{x}_1$	0.1	0.0	0.0	0.0	Ham
$\tilde{\mathbf{x}}_1$	0.1	0.0	**0.3**	**0.453**	**Spam**

Table: Examples of original samples $\boldsymbol{x}_i$ and contrastive samples $\tilde{\boldsymbol{x}}_i$

1. Constraint on the contrastive class:

$$argmax(f(\tilde{x})) \neq argmax(f(\boldsymbol{x})) \tag{1}$$

2. Constraint on the # of key features:

$$|\mathcal{S}| \leq K \tag{2}$$

3. Constraint on the mutual information:

$$\mathrm{SU}(\mathcal{X}^i, \mathcal{X}^j) \leq \gamma \quad \forall i, j \in \mathcal{S} \tag{3}$$

4. Constraint on the domain:

$$\tilde{\mathbf{x}} \in dom(\mathcal{X}) \tag{4}$$

Objective Function

Objective Function

Given $\boldsymbol{x}$, hyperparameter K, γ, our goal is to generate new contrastive sample $\tilde{\mathbf{x}}$ to explain the prediction $f(\boldsymbol{x})$ by solving the objective function:

$$\begin{aligned} \min_{\tilde{\boldsymbol{x}}} \quad & dist(\tilde{\boldsymbol{x}}, \boldsymbol{x}) \\ s.t. \quad & \mathrm{argmax}(f(\boldsymbol{x})) \neq \mathrm{argmax}(f(\tilde{\boldsymbol{x}})), \quad |\mathcal{S}| \leq K \\ & \mathrm{SU}(\mathcal{X}^i, \mathcal{X}^j) \leq \gamma \quad \forall i, j \in \mathcal{S}, \quad \tilde{\mathbf{x}} \in dom(\mathcal{X}) \end{aligned} \tag{5}$$

Generation Algorithm

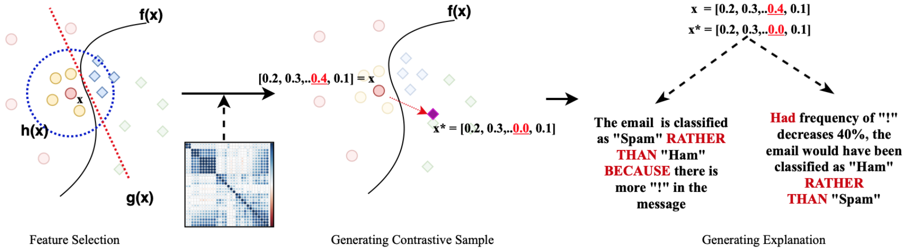

Figure: Grace Algorithm - Local-Based Feature Selection

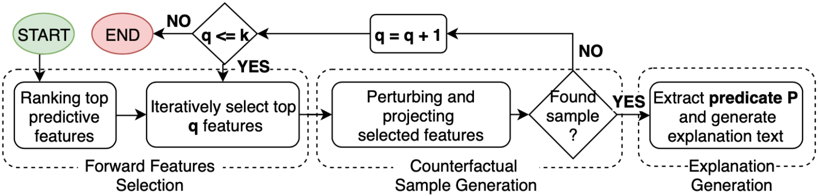

Figure: Grace Algorithm

OUTLINE

Experiment - Dataset & Baseline

Table: Dataset statistics and prediction performance

Dataset	#Class	#Feat.	#Data	Acc.*	F1*
eegeye	2	14	14980	0.858	0.858
diabetes	2	8	768	0.779	0.777
cancer95	2	9	699	0.963	0.963
phoneme	2	5	5404	0.774	0.772
segment	7	19	2310	0.836	0.817
magic	2	10	19020	0.862	0.859
biodeg	2	41	1055	0.853	0.851
spam	2	57	4601	0.932	0.932
cancer92	2	30	569	0.958	0.958
mfeat	10	216	2000	0.943	0.936
musk	2	166	476	0.783	0.789

(*) Accuracy and F1 scores are averaged across 10 different runs.

1. **NearestCT**: Select nearest contrastive sample from the training set
2. **DeepFool** [Moosavi-Dezfooli et al., 2016]: Generate adversarial samples with $\min_{\tilde{x}} \|\tilde{x} - \boldsymbol{x}\|_2$
3. **Lime** [Ribeiro et al., 2016]: Instance-based explanation method

Experiment - Quantitative

1. Conciseness:

$$\mathbf{R}_{\text{fidelity}} = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in \tilde{\mathcal{X}}} \mathbb{1}(\tilde{\mathbf{y}} == \text{argmax}(f(\tilde{\mathbf{x}}))) \tag{6}$$

$$\mathbf{R}_{\text{avg\#Feats}} = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} |\mathcal{S}_{\tilde{\mathbf{x}}}| \tag{7}$$

2. Info-Gain:

$$\mathbf{R}_{\text{info-gain}} = 1 - \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \sum_{i \in \mathcal{S}_{\tilde{\mathbf{x}}}} \sum_{j \in \mathcal{S}_{\tilde{\mathbf{x}}}} \frac{\mathrm{SU}\left(\mathcal{X}^i, \mathcal{X}^j\right)}{|\mathcal{S}_{\tilde{\mathbf{x}}}|^2} \tag{8}$$

3. Influence:

$$\mathbf{R}_{\text{domain}} = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\tilde{\boldsymbol{x}} \in \tilde{\mathcal{X}}} \mathbb{1}(\tilde{\boldsymbol{x}} \in \text{dom}(\mathcal{X})) \tag{9}$$

$$\mathbf{R}_{\text{influence}} = \frac{\mathbf{R}_{\text{fidelity}} \times \mathbf{R}_{\text{info-gain}} \times \mathbf{R}_{\text{domain}}}{\mathbf{R}_{\text{avg\#Feats}}} \tag{10}$$

Experiment - Quantitative Results

		biodeg	spam	cancer92	mfeat	musk
$\mathbf{R}_{\text{avg\#Feats}}$	NearestCT	20.53	17.50	29.97	204.22	147.86
	DeepFool	41.00	57.00	30.00	216.00	166.00
	GRACE-Local	<u>3.07</u>	<u>2.95</u>	**3.95**	<u>3.28</u>	<u>3.74</u>
	GRACE-Gradient	**1.93**	**1.09**	<u>4.5</u>	**2.76**	**2.85**
$\mathbf{R}^{*}_{\text{info-gain}}$	NearestCT	0.44	<u>0.62</u>	0.02	<u>0.58</u>	0.28
	DeepFool	<u>0.58</u>	0.53	0.01	**0.59**	0.29
	GRACE-Local	0.46	0.47	**0.13**	0.34	<u>0.3</u>
	GRACE-Gradient	**0.76**	**0.95**	<u>0.04</u>	0.50	**0.4**
$\mathbf{R}_{\text{influence}}$	NearestCT	0.02	0.04	0.00	0.00	0.00
	DeepFool	0.01	0.01	0.00	0.00	0.00
	GRACE-Local	<u>0.15</u>	<u>0.16</u>	**0.04**	<u>0.1</u>	<u>0.08</u>
	GRACE-Gradient	**0.4**	**0.88**	<u>0.01</u>	**0.18**	**0.14**

Table: All results are averaged across 10 different runs. The best and second best results are highlighted in **bold** and <u>underline</u>.

Experiment - Quantitative - Conciseness

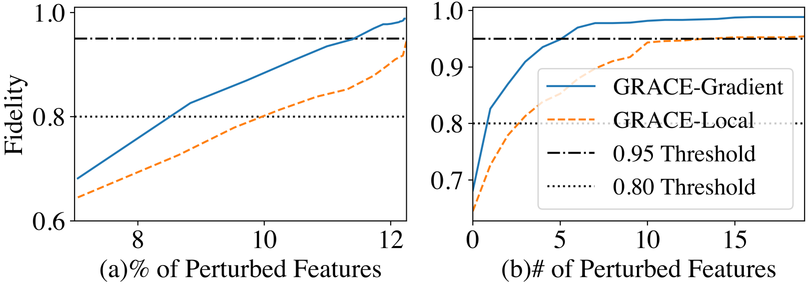

Figure: Percentage of Perturbed Features v.s. Fidelity

Experiment - User-Study

1. Comparison between Grace and Lime
   - Hypothesis $\mathcal{H}_1$: **more intuitive and friendly**
   - Hypothesis $\mathcal{H}_2$: **more comprehensible**
   - Hypothesis $\mathcal{H}_3$: **leads to better post-explanation decisions**
2. Recruit Amazon MTurk workers as general end-users
3. No assumptions on workers' prior knowledge in machine learning

Experiment - User-Study

Feature	Bare_nuclei	MarAdh	CluThic	mitoses	CelSizUni	CelShaUni	NorNuc	SinEpCeSi	**Model Prediction**
Value	10.0	5.0	8.0	1.0	8.0	5.0	3.0	2.0	**Malignant**

Explanation

*"Had **bare_nuclei** been 3.0 point lower and **CluThic** been 7.0 point lower, the patient would have been diagnosed as Benign rather than Malignant"*

Q1:Given a scale from 1 to 10, "how intuitive and friendly is the explanation to you?" (1 is least preferable, 10 is most preferable)

0

Q2:Given a scale from 1 to 10, "how understandable is the explanation to you?" (1 is least preferable, 10 is most preferable)

0

Figure: Example of an User-study Task for $\mathcal{H}_1$, $\mathcal{H}_2$

Experiment - User-Study

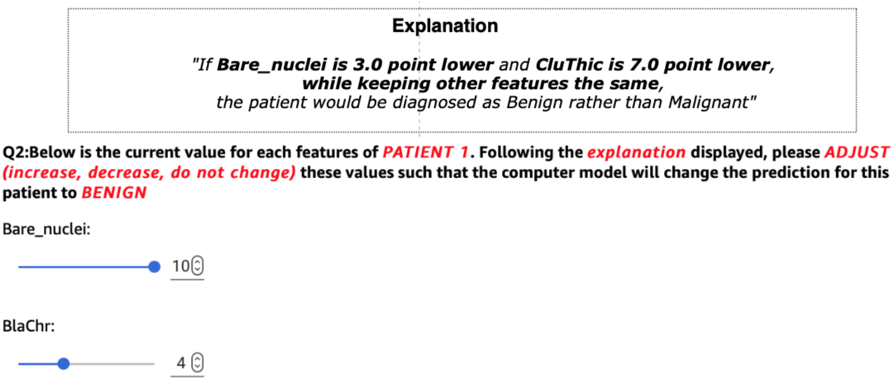

Figure: Example of an User-study Task for $\mathcal{H}_3$

Experiment - User-Study

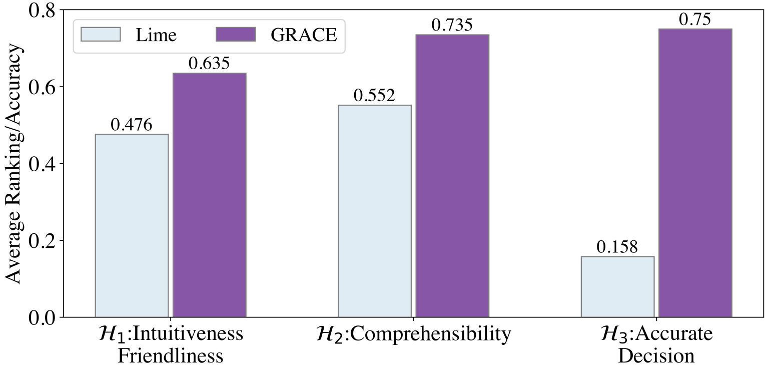

Figure: Comparison of generated explanation: GRACE v.s. Lime. Scores are normalized to [0,1]. All results are statistically significant ($\mathcal{H}_1 : p - value < 0.05$, $\mathcal{H}_2, \mathcal{H}_3 : p - value < 0.01$)

Sensitivity Analysis - K, γ

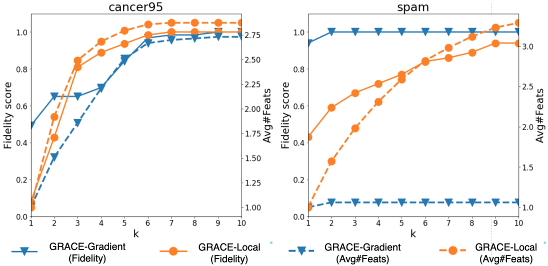

Figure: Sensitivity of K on Fidelity

Table: Effects of entropy threshold γ on Info-Gain

Dataset	Method	1.0	0.7	0.5	0.3
musk	GRACE-Gradient	0.51	0.51	**0.58**	**0.58**
	GRACE-Local	0.36	0.36	**0.54**	**0.54**
segment	GRACE-Gradient	0.57	0.57	**0.59**	**0.59**
	GRACE-Local	0.79	0.79	**0.84**	**0.84**

OUTLINE

1. ✓ Introduction
2. ✓ Motivation & Challenges
3. ✓ Explanation by Intervention
4. ✓ GRACE Algorithm
5. ✓ Experiments
6. Conclusion

Future Direction & Conclusion

1. **GRACE**: A novel instance-based algorithm that provides end-users with simple natural text explaining neural network models' predictions in a contrastive *"Why X rather than Y"* fashion.
2. **GRACE**: more intuitive, friendly, comprehensible and leads to more accurate decisions than Lime

Additional Information

1. Source Code and Slides:
   `https://github.com/lethaiq/GRACE_KDD20`
2. Pike Group at Penn State:
   `http://pike.psu.edu`

References I

[Moosavi-Dezfooli et al., 2016] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016).
Deepfool: a simple and accurate method to fool deep neural networks.
In *Proceedings of the 2016 IEEE CVPR*, pages 2574–2582.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
"why should I trust you?": Explaining the predictions of any classifier.
In *KDD*, pages 1135–1144.