THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Indian Institute of Technology
Roorkee, India
भारतीय प्रौद्योगिकी संस्थान
रुड़की, भारत

Thapar Summer School 2022 | Day 27

'Explainable AI'
Introduction and Hands-on Practices

Puneet Kumar
puneet-kr.github.io

Research Scholar
Machine Intelligence Lab, CSE Dept., IIT Roorkee

Founding Director & CTO
PaiByTwo Private Limited

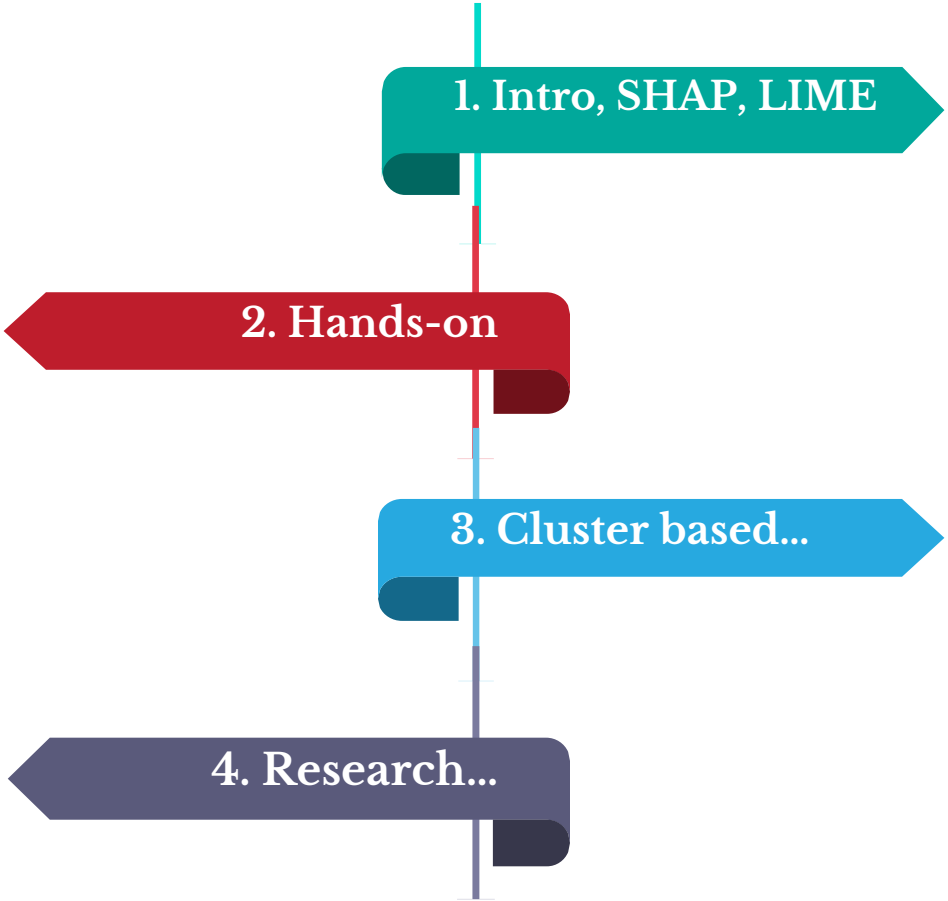
PaiByTwo

Computer Science & Engineering

About the Speaker

- Founding Director and CTO, PaiByTwo Pvt. Ltd. [2021+]
- Visiting Researcher, Osaka Prefecture University, Osaka, Japan [2019]
- Visiting Researcher, Samsung R & D, Delhi, India [2018-19]
- Software Engineer, Oracle IDC, Hyderabad, India [2014-16]
- Ph.D., IIT Roorkee, Uttarakhand [2018-22]
- M.E. (CSE), TIET Patiala, Punjab [2016-18]
- B.E. (CSE), MIT Manipal, Karnataka [2010-14]
- Areas of Research
  - Affective Computing and Cognitive Science
  - Multimodal Intangible Emotion Analysis
  - Machine Learning and Deep Learning
  - Explainable Artificial Intelligence
  - Meta-heuristic Optimization

Outline



1. Explainability
2. Interpretability
3. SHAP
4. LIME

1. SHAP
2. LIME
3. Interaction

1. Cluster-based Explainability
2. Transfer Learning recap
3. Hands-on

1. Extend Shap/LIME
2. DnCShap
3. Research papers
4. Resources
5. Q/A

Explainability

Interpretability

Introduction

Fairness

Trustworthy AI

Reliability

Explainability & Interpretability [1]

- **Explainability**:
  - Describe a model's mechanism that led to a particular output.
  - **How** did we reach a particular output?

- **Interpretability**:
  - Understand the context of a model's output, analyzes its functional design, and relate the design to the output.
  - **What** made us reach a particular output?
  - Cause-effect relationship. '**What**' was the cause of the output effect?

Local & Global Explainability

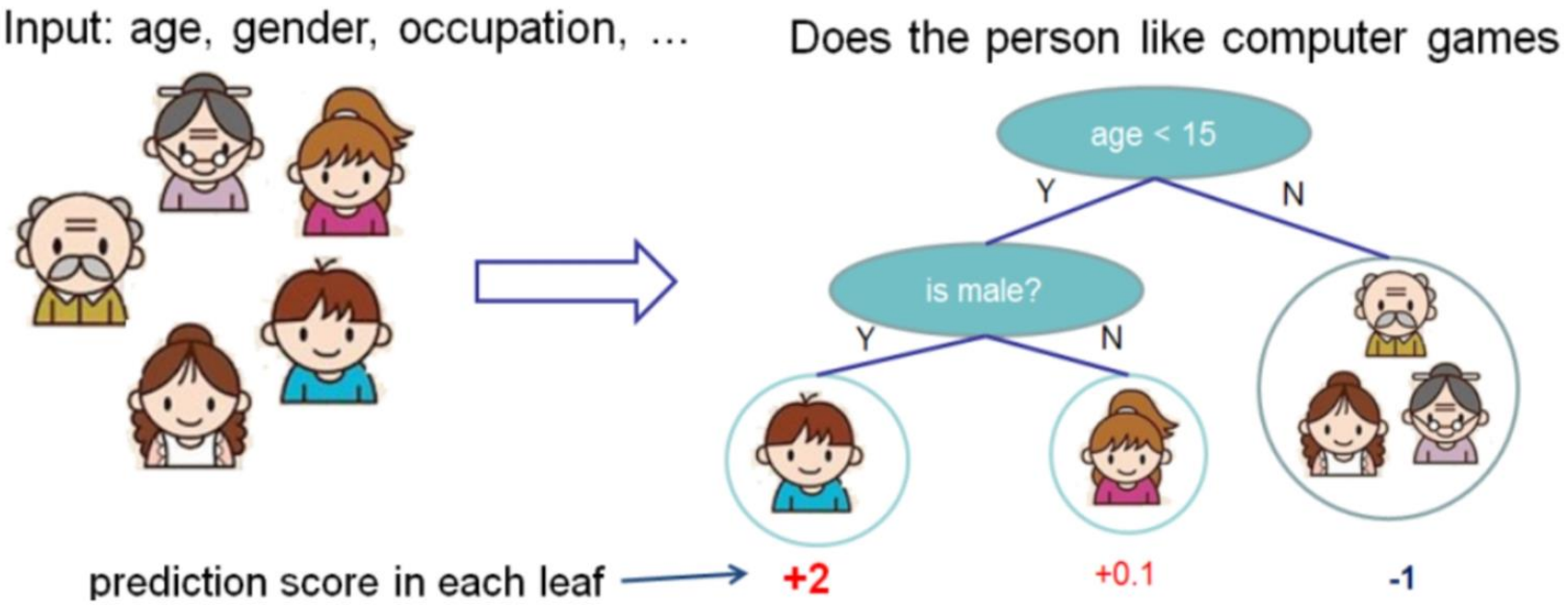


Figure 1: Local & Global Explainability

Explainable AI Approaches

- Explainable AI approaches:
  - **Attribution-based approaches**, such as **SHAP** [2] assign a relevance or importance score to each input feature based on Shapely values [3].
  - **Perturbation-based approaches**, such as **LIME** [4] compute the importance score by slightly changing the input.
  - **Backpropagation-based approaches**, such as 'Saliency Map' [6] and Grad-CAM [5] calculate the attributions by back-propagating through the network.

Popular Explainability Models

Popular Explainability Models

- Popular Explainability models
  - **SHAP**: Attribution based.
  - **LIME**: Perturbation based.
  - Grad-CAM: Back-propagation based. [self-study]

- These are **Surrogate Models**.
  - They still use the black-box machine learning models.
  - They tweak the input slightly and test the changes in prediction.
  - This tweak has to be small so that it is still close to the original data.

SHAP

‘SHapley Additive exPlanations.’

- Attribution based Explainability approach.
- Based on Game Theory (reward proportional to contribution).
- Break (virtually) the core complex model into many simple models. Find Shapely values example-by-example.
- Then Aggregate and Explain the overall model.

SHAP

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

Why bother explaining ?

- Simple linear regression model (coefficient Ø for feature **x**)

$$f(x_1, x_2, \ldots, x_n) = \phi_1 x_1 + \phi_2 x_2 + \ldots + \phi_n x_n$$

- It can discover linear relationships. It is explainable.
- Non-linear relationships <= complex models, difficult to explain.
- What do we want?
  - To be able to use complex models.
  - Learn non-linear relationships in the data.
  - To be able to interpret them as well.

How to (start) explaining ?

- One (linear regression) model: explain one example.
- Many linear regression models: explain all the examples.
- Rephrasing:
  - "Take a complex model, which has learnt non-linear patterns in the data, and broken it down into lots of linear models which describe individual data points."
- This is how SHAP does it ! [More details: upcoming slides.]
- Idea: Breaking into parts | Complexity | Divide & Conquer.
- Hint: This is one way to develop novel research ideas.

Shapely Values

- To explain how the complex model behaved for just one data point.
- Aggregate it to get an idea of how my model worked globally.
- Assume a linear explanation model, **g**

$$g_{Frank} = \phi_{FrankAge} + \phi_{FrankGender} + \phi_{FrankJob}$$

- Shapley value, ϕ denotes the importance of a particular feature.
- $\phi_i(p)$ for a certain feature **i** (out of **n** total features), given a prediction **p** (by the complex model) is given as follows, where **S** denotes a set.

$$\phi_i(p) = \sum_{S \subseteq N/i} \frac{|S|!(n - |S| - 1)!}{n!}(p(S \cup i) - p(S))$$

Shapley values: Elaboration

- Shapely value: inherited from Game Theory [3].
- Core Idea: Contribution ∝ Reward.
- Shapley values: calculate the importance of a feature by comparing what a model predicts with and without the feature.
  - Calculate what the prediction of the model would be without feature **i**,
  - Calculate the prediction of the model with feature **i**,
  - Calculate the difference => importance of feature **i**.

$$\text{Importance of i} = p(\text{with i}) - p(\text{without i})$$

- Example calculations of ϕ : SHAP base paper [2] and blog [7].

LIME

'**Local** Interpretable **Model-agnostic** Explanations.'

- Purturbation based Explainability approach.
- Algorithm to explain **any** *black-box algorithm by making **Local** approximations.
- It doesn't know the internals of the blak-box => **Model-agnostic**.
- **Core idea**: zoom into the local area of the individual prediction => make a simple explanation in that local region.

---

*Most complex models are complete black-boxes and their internal mechanisms are hidden.

LIME

"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.

LIME: Core Idea

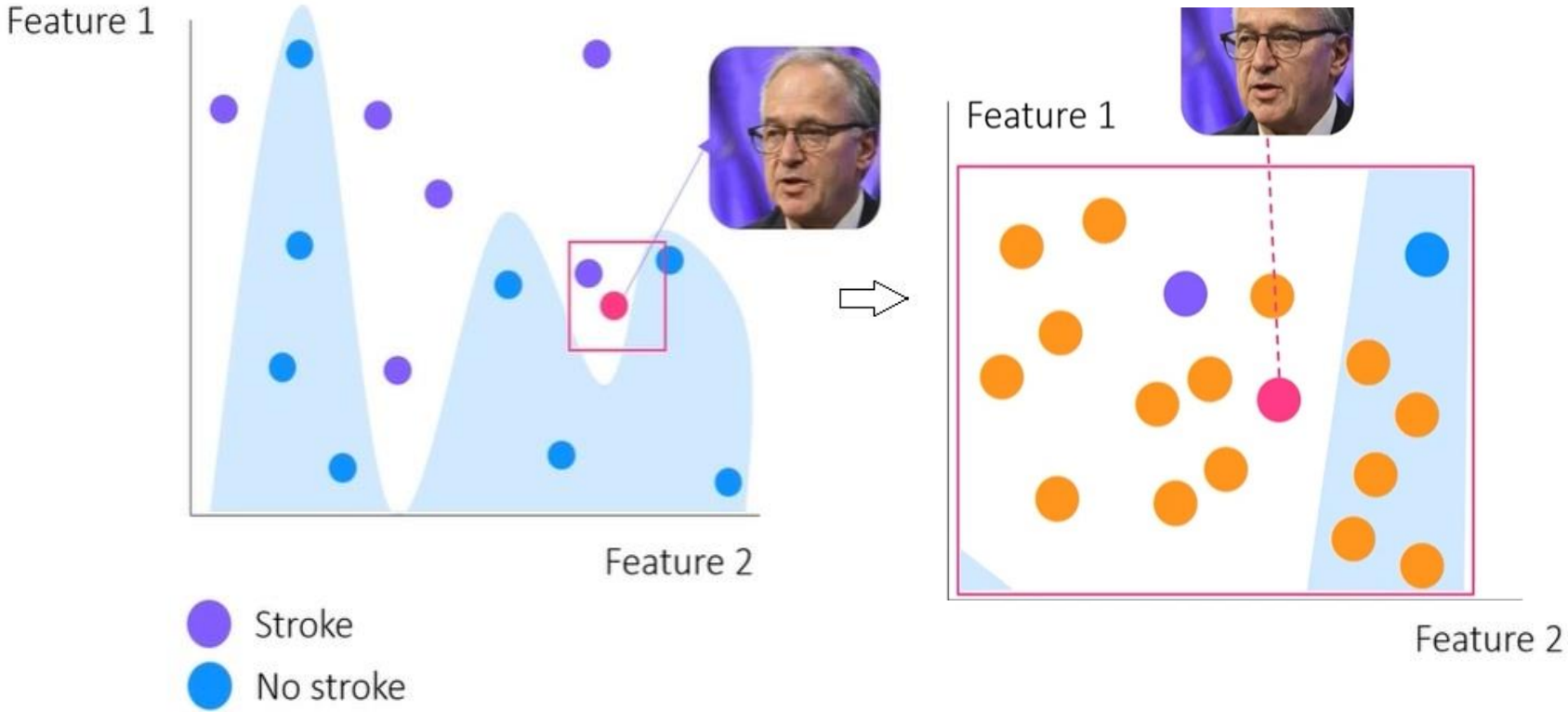


Figure 2: Code Idea of LIME

LIME: The Maths

Local approximation of our complex model for a specific input.

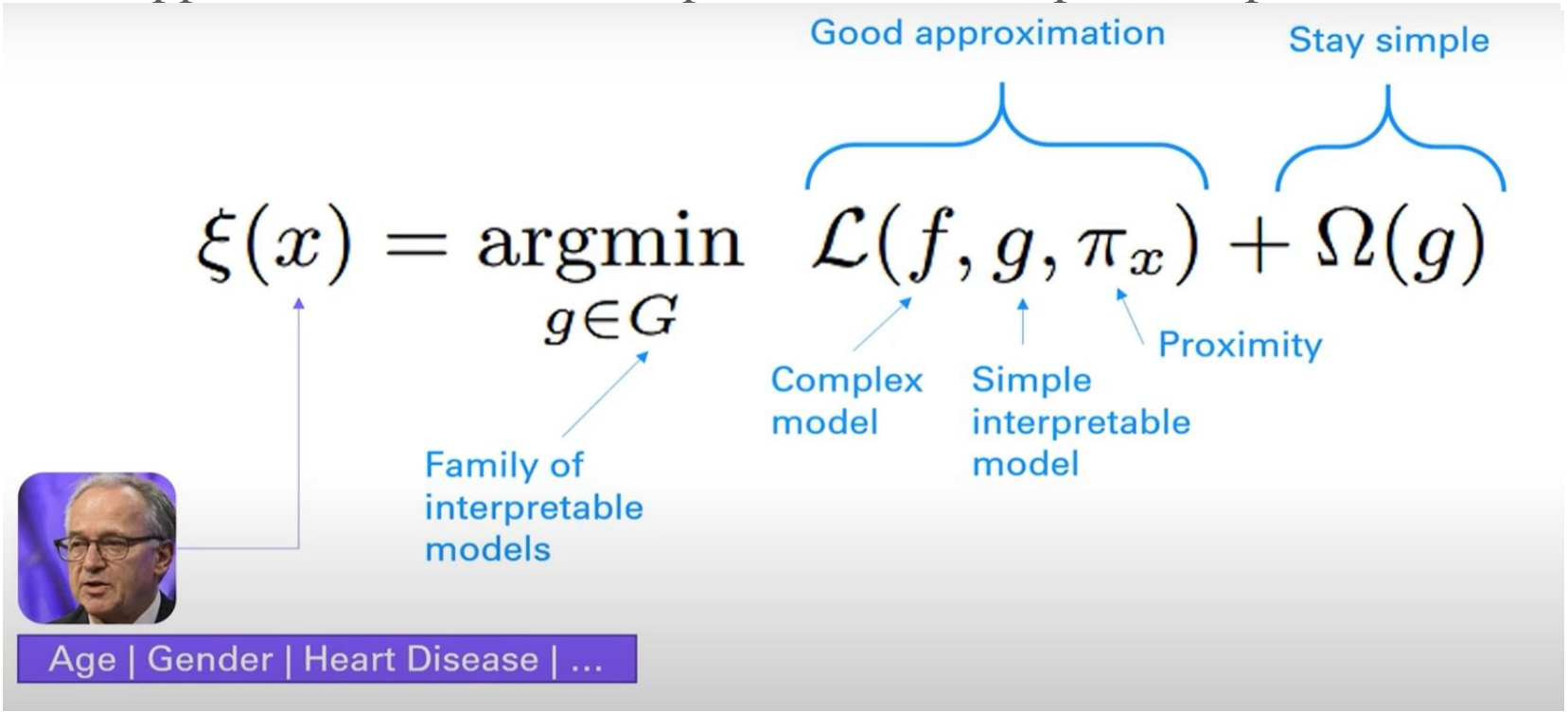


Figure 3: The mathematics behind LIME

LIME: Elaboration

- Purturb (e.g. add/remove/transform some binary dimensions).
- => generate new data in the proximity of the target data sample.

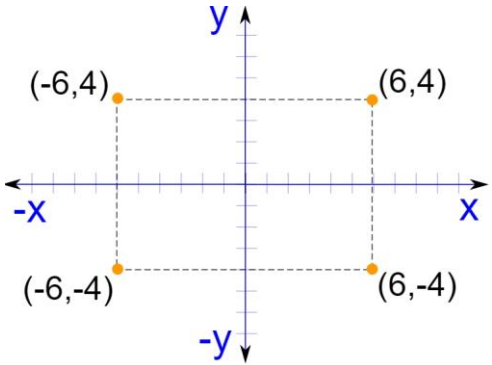


Figure 4: Purturbation's example

- Feed the purturbed data to the black-box model.
- Observe the changes in the outputs => make sense of it => explain.

Hands-on

Ref: https://github.com/puneet-kr/TSS2022D27

Hands-on: LIME

☑ Import libraries, setup the model and load the data

☑ Install LIME (pip install lime)

☑ NLP Pre-processing

☑ Split train-test data and create the model

Note: Perturbation using scikit-learn's *make_pipeline* algorithm that vectorizes and transforms the data

☑ Create Model Explainer (LIME_explainer)

☑ Explain using LIME_explainer

Hands-on: SHAP

☑ Import libraries, setup the model and load the data

☑ Install SHAP (pip install shap)

☑ Pre-processing

☑ Split train-test data and create the model

☑ Create SHAP_explainer and calculate the shap values

Note: Attribution/perturbation using scikit-learn's make_pipeline algorithm that vectorizes and transforms the data

☑ Interpret Global Feature Importance

☑ Interpret Local Feature Importance

Cluster based Explainability

Cluster based Explainability

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Towards the explainability of Multimodal Speech Emotion Recognition

Puneet Kumar[†]*, *Vishesh Kaushik*[‡]*, *Balasubramanian Raman*[†]

[†]Computer Science and Engg. Dept., Indian Institute of Technology, Roorkee, India, 247667
[‡]Mechanical Engg. Dept., Indian Institute of Technology, Kanpur, India, 208016

pkumar99@cs.iitr.ac.in, kvishesh@iitk.ac.in, bala@cs.iitr.ac.in

Abstract

In this paper, a multimodal speech emotion recognition system has been developed, and a novel technique to explain its predictions has been proposed. The audio and textual features are extracted separately using attention-based Gated Recurrent Unit (GRU) and pre-trained Bidirectional Encoder Representations from Transformers (BERT), respectively. Then they are concatenated and used to predict the final emotion class. The weighted and unweighted emotion recognition accuracy of 71.7% and 75.0% has been achieved on Emotional Dyadic Motion Capture (IEMOCAP) dataset containing speech utterances and corresponding text transcripts. The training and predictions of network layers have been analyzed qualitatively through emotion embedding plots and quantitatively by analyzing the intersection matrices for various emotion classes' embeddings.
Index Terms: Multimodal emotion recognition, deep network explainability, intersection matrix, embedding plot.

Explaining the internal mechanism of deep-learning-based classification methods has emerged as a recent research topic [21]. In this context, Lin et al. [22] proposed a method for interpreting multimodal emotion recognition of biological signals using deep-learning. In another work, Zhang et al. [23] analyzed the effect of various audio features on emotion arousal and interpreted the corresponding response. Riberio et al. [24] developed a technique to find the input's part responsible for a particular output. In another work, Shrikumar et al. [25] came up with a method to break down the output predictions by tracing the contributions of all the neurons. However, the above-discussed methods were unable to show the network's layer-by-layer training. It inspired us to develop a method to explain and interpret a DNN based SER system's predictions.

The proposed system encompasses a speech emotion recognition (SER) module and a text emotion recognition (TER) module. For the SER module, Gated Recurrent Units (GRUs) [26] have been implemented along with attention to ex-

Cluster based Explainability

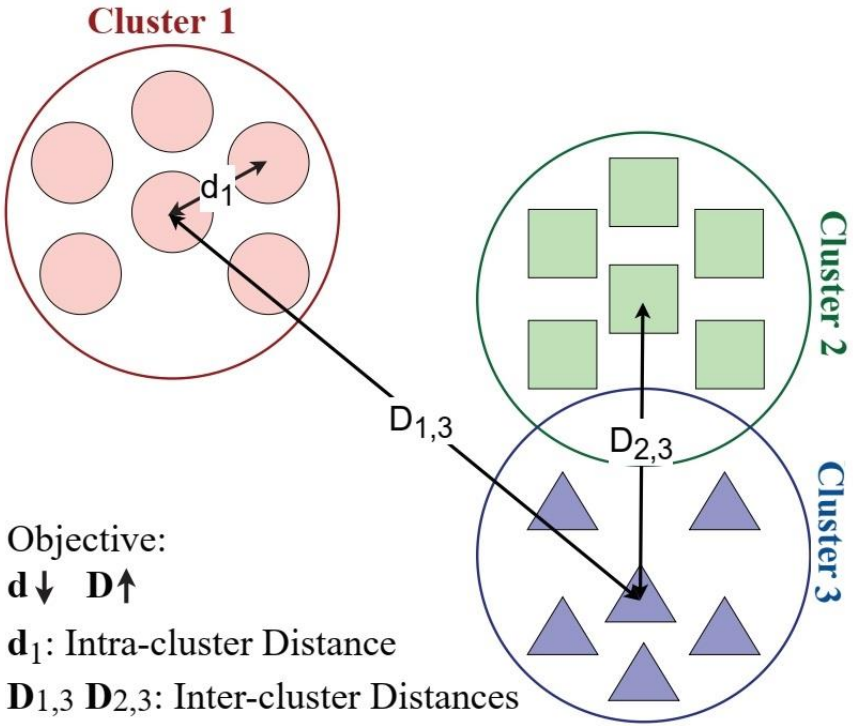


Figure 5: Core hypothesis of Cluster-based Explainability

Cluster based Explainability

Proposed Explainability technique [9] to explain the trained model's predictions:

➵ Compute principal components (x_i, y_i, z_i) of the last layer's embedding for class i.

➵ For class i and (x_i, y_i, z_i) components, compute mean m_i and standard deviation σ_{m^-i}.

➵ For i^{th} class's m^{th} component's spread, compute the range for i^{th} motion with left extreme point $L_i(m)$ and right extreme point $R_i(m)$.

➵ The intersection between classes i and j, *i.e.*, $I_{i,j}(m)$ is computed as the intersection between the spread of the m^{th} component's data for classes i and j.

➵ The total intersection between classes i and j, i.e., $I_{i,j}$ is calculated as the product of component-wise intersections $I_{i,j}(x) * I_{i,j}(y) * I_{i,j}(z)$.

➵ The $(i, j)^{th}$ element of the *Intersection Matrix* I of size $c \times c$, represents the total intersection $I_{i,j}$ between the emotion classes i & j where c denotes total no. of classes.

Cluster based Explainability

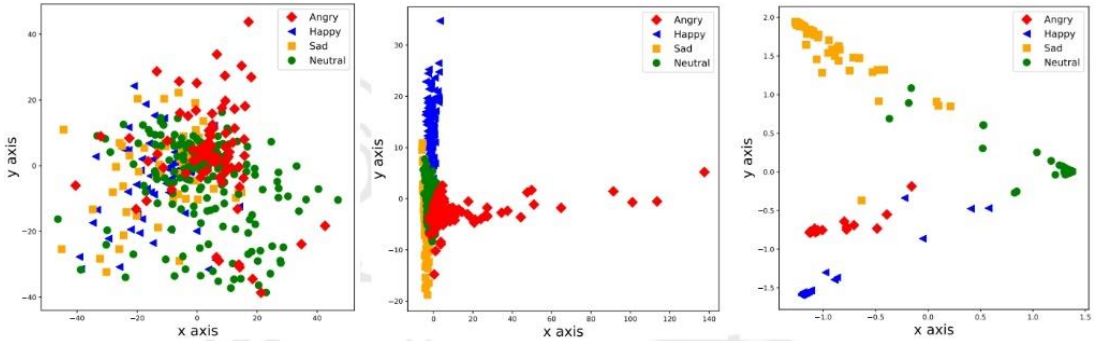


Figure 6: Emotion embedding plots for various layers

Table 1: Intersection matrices for the proposed model.

	Third-last Layer				Second-last Layer				Output Layer			
	Angry	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*	*Angry*	*Happy*	*Sad*	*Neutral*
Angry	1.0000	0.0023	0.0014	0.0042	1.0000	0.0022	0.0047	0.0049	1.0000	0.0000	0.0000	0.0000
Happy		1.0000	0.0399	0.0453		1.0000	0.0242	0.0458		1.0000	0.0000	0.0000
Sad			1.0000	0.0280			1.0000	0.0179			1.0000	0.0001
Neutral				1.0000				1.0000				1.0000

Hands-on

Ref: https://github.com/puneet-kr/TSS2022D27

Hands-on: Cluster based Explainability

- ☑ Import Libraries
- ☑ Create Transfer Learning pipeline (or setup a new model)
- ☑ [ToDo]: Add/reduce more layers and observe the clusters
- ☑ Setup Data Generators
- ☑ Define step-size and for the model
- ☑ Cluster plot implementation
- ☑ Model Summary
- ☑ [ToDo] Cluster plots for various layers
- ☐ Intersection Matrix: Ref. https://github.com/MIntelligence-Group/SpeechImg_EmoRec

Explainable AI Research

DnCSHAP [10]

'Divide and Conquer based SHapley Additive exPlanations.'

- Computes the approximated Shapley values for the input images in linear time instead of the exponential.
- Divide and Conquer => continuously divide the input image into two parts. Find Shapely values. Aggregate.
- Identifies the highly relevant parts of the input that contribute the most to recognizing the emotion classes.

Explainable Facial Emotion Recognition

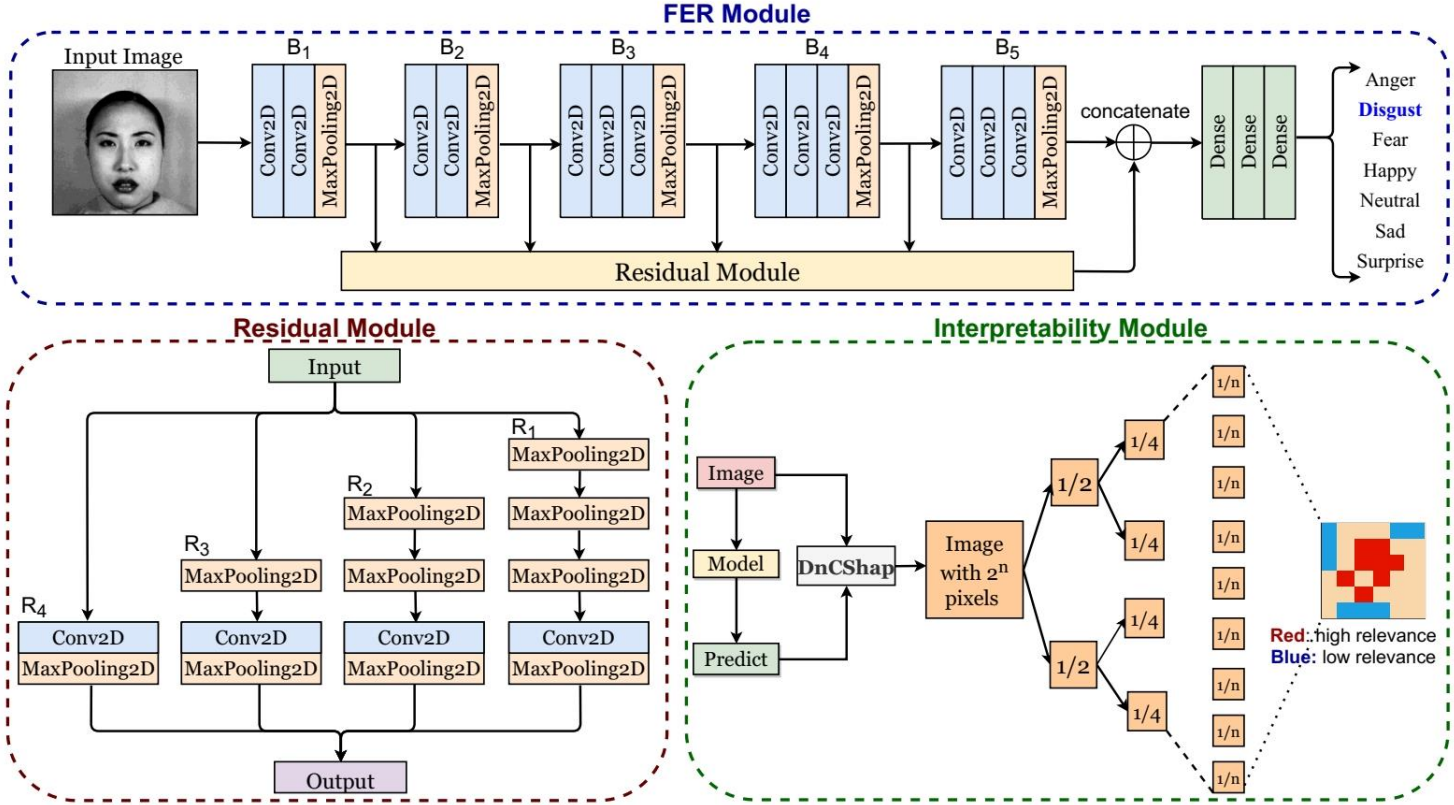


Figure 7: Extension of SHAP model to interpret facial emotion recognition

Explainable Facial Emotion Recognition

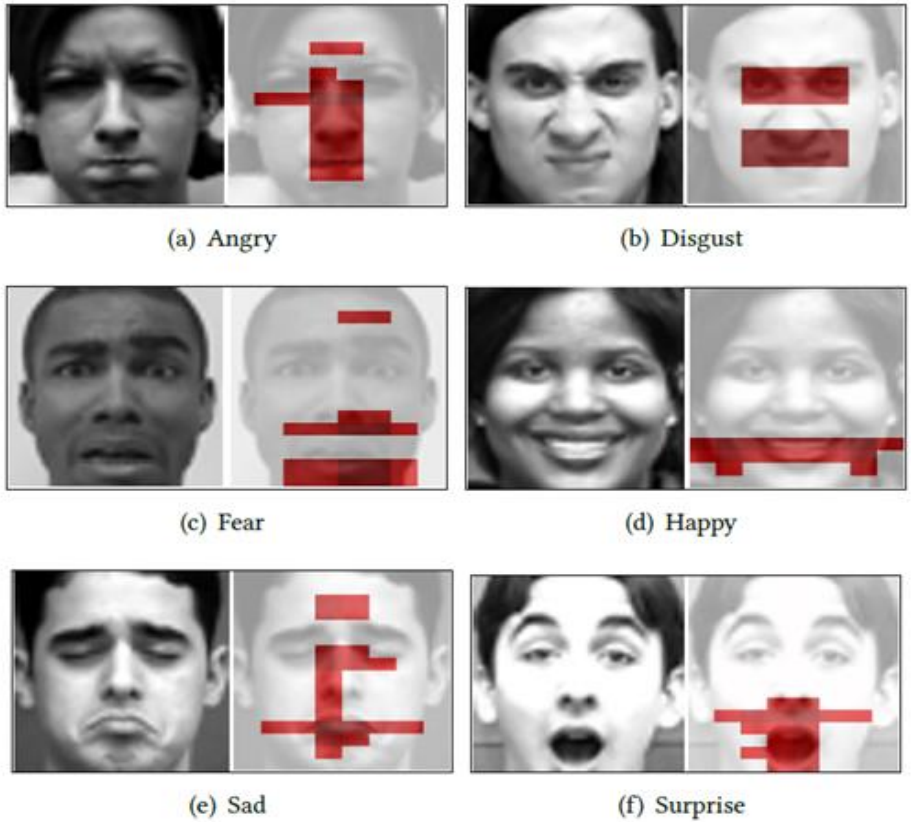


Figure 8: Results of DnCSHAP for facial emotion recognition

Explainable Multimodal Emotion Recognition*

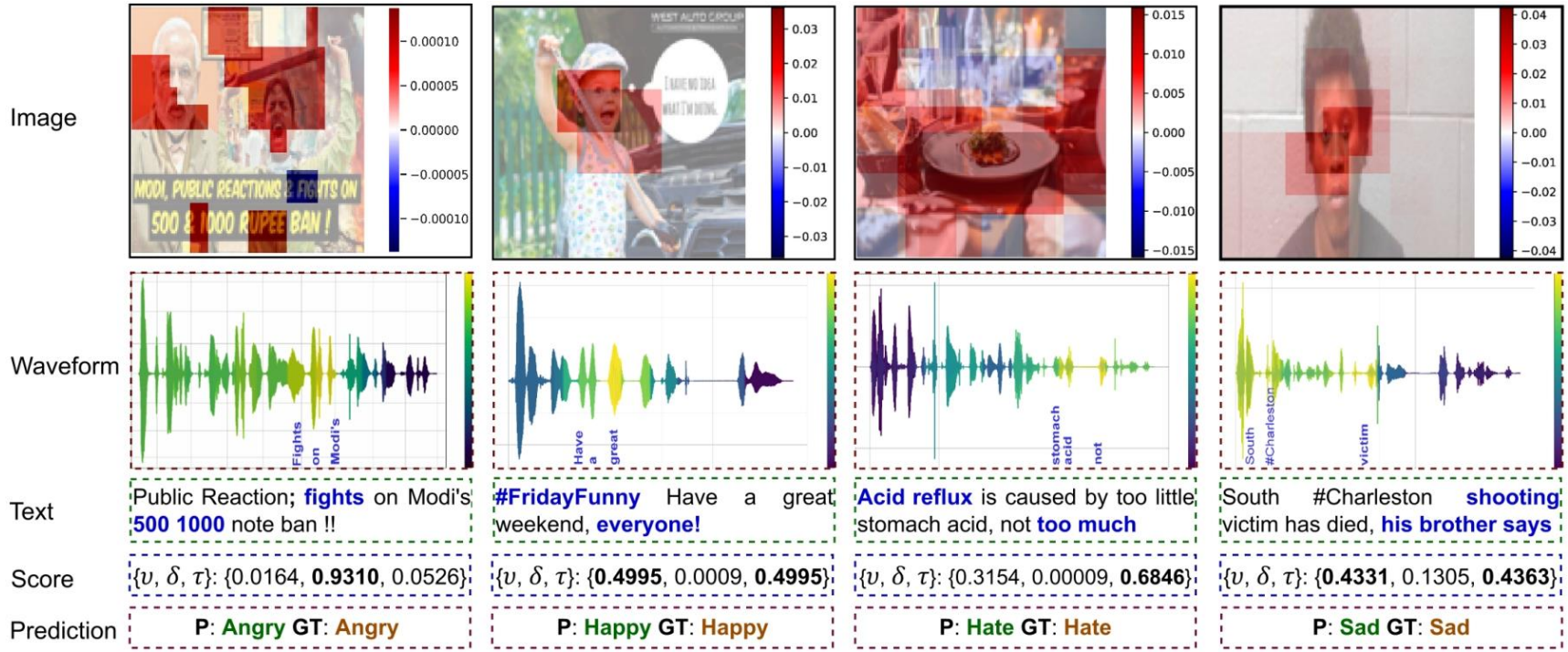


Figure 9: Results of DnCShap for multimodal emotion recognition

*Under review work. Ref: https://github.com/MIntelligence-Group/SpeechImg_EmoRec

Resources

Papers, Tutorials, Blogs & Talks

- SHAP Paper
- SHAP Documentation
- SHAP Implementation
- SHAP Results' Analysis
- LIME Paper
- LIME Documentation
- LIME Implementation (inc. Results' Analysis)
- Blog: Interpretable ML
- Book: Interpretable ML
- Talk: building Explainable ML Systems: The Good, Bad, and Ugly
- Tutorials: Explainable AI | Playlist (DeepFindr)

Advanced Concepts

- Grad-CAM : Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. CVPR 2017 (pp. 618-626).

- DnCSHAP: Malik, S., Kumar, P. and Raman, B., Towards Interpretable Facial Emotion Recognition. ICVGIP 2021 (pp. 1-9).

- Multimodal Interpretability Example: github.com/MIntelligence-Group/SpeechImg_EmoRec

Key References

1. David A Broniatowski et al. Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST), 2021
2. Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In The 31st International Conference on Neural Information Processing Systems (NeuroIPS). 4768–4777.
3. LS Shapley. 1953. A Value for n-person Games, Contributions to the Theory of Games II, AW Tucker, HW Kuhn.
4. Marco Ribeiro et al. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 1135–1144.
5. Ramprasaath Selvaraju et al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.
6. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034 (2013).
7. https://medium.com/@gabrieltseng/interpreting-complex-models-with-shap-values-1c187db6ec83
8. https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b
9. Kumar, P., Kaushik, V., & Raman, B.. Towards the Explainability of Multimodal Speech Emotion Recognition. Interspeech (pp. 1748-1752).
10. Malik, S., Kumar, P. and Raman, B., Towards Interpretable Facial Emotion Recognition. ICVGIP 2021 (pp. 1-9).

Thank you

puneet-kr.github.io

pkumar99@cs.iitr.ac.in

classroom@paibytwo.com