Paper Review

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability

NeurIIPS 2021

Dylan Slack[1] , Sophie Hilgard[2], Sameer Singh[1] , Himabindu Lakkaraju[2]
[1]UC Irvine [2]Harvard University

Min-Seok Yang
Natural Language Processing Lab
Department of Artificial Intelligence, Kyung Hee University

Part 1.

Background

- **Explain machine learning model in deployment**
  - Model depolyment in domains such as healthcare, medicine, law, finance
    - Important to ensure that decision makers have a clear understanding of the behavior of these models
  - Explain complex black box models by constructing interpretable local approximations
    - Post-hoc explanations construct interpretable local approximations
    - Lime, SHAP, MAPLE, Anchors

Post-hoc interpretability

- Interpretability is achieved by applying methods that analyze the model after training

Local Method

- Model-agnostic methods which focus on explaining individual prediction

Christoph Molnar. Interpretable Machine Learning (Third Edition) . https://christophm.github.io/interpretable-ml-book/

Part 1.

Background

- **Existing local explanation methods**
  - Lime, SHAP, MAPLE, Anchors
  - Unstablility
    - Negligibly small perturbations to an instance can result in substantially different explanation
  - Inconsistency
    - Multiple runs on the same input instance with the same parameter settings may result in vastly different explanations

Background

- **Reliable metrics to ascertain the quality of the explanations**
  - Explanation fidelity rely heavily on the implementation details of explanation method
    - No guidance on determining the values of certain hyperparameters that are critical to the quality of the resulting local explanations (e.g., number of perturbations in case of LIME)
  - Local explanation methods are also computationally inefficient
    - Typically require a large number of black box model queries to construct local approximations

Explanation produced by LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Minimize $\boxed{\mathcal{L}(f, g, \pi_x)}$ while having $\boxed{\Omega(g)}$ be low enough to be interpretable by humans

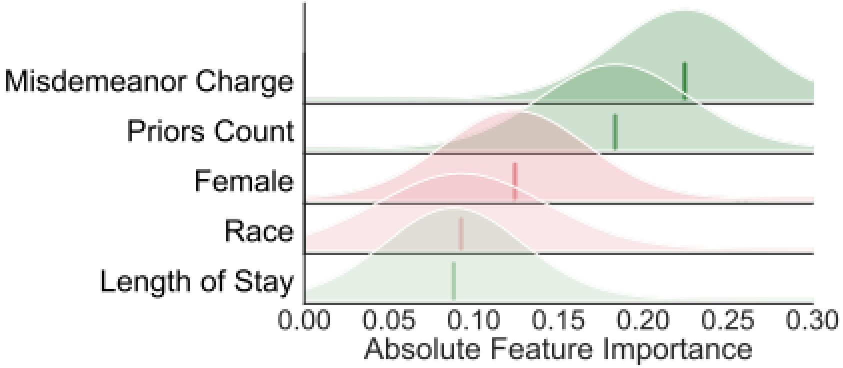
Formulation

- Explanation family G
- Fidelity function $\mathcal{L}$
- Complexity measure Ω

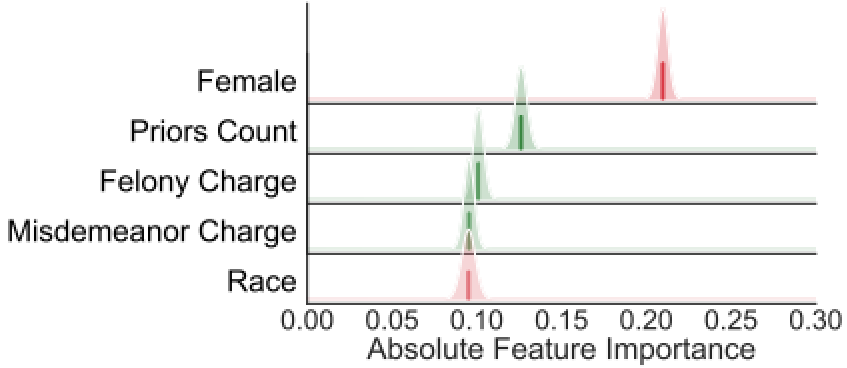
Marco Tulio Ribeiro et al. Why Should I Trust You? explaining the predictions of any classifier. KDD. 2016.

Introduction

- **Bayesian framework for generating local explanations with uncertainty**
  - Generate consistent, stable, and reliable explanations with guarantees in a computationally efficient manner
  - 1. Bayeslime & BayesSHAP: Bayesian versions of LIME and KernelSHAP



(a) Explanation computed with 100 perturbations



(b) Explanation with 2000 perturbations

Example Explanation

- Dataset: COMPAS
- Vertical Lines: LIME
- Shaded Region: BayesLIME
- Red: Negative Effect
- Green: Positive Effect

LIME: Contradictory feature importance for different number of perturbations

BayesLIME: Proveide more context (i.e., tighter uncertainty interval indicates importance)

Introduction

- **Bayesian framework for generating local explanations with uncertainty**
  - Generate consistent, stable, and reliable explanations with guarantees in a computationally efficient manner
  - 1. Bayeslime & BayesSHAP: Bayesian versions of LIME and KernelSHAP
  - 2. Closed form expressions for the posteriors of the explanations
    - Eliminate the need for any additional computational complexity
  - 3. Credible intervals produced by our framework
    - Make concrete inferences about the quality of the resulting explanations
    - Produce explanations that satisfy user specified levels of uncertainty (e.g., 95% confidence level)

Method

- **Notation**
  - $f : \mathbb{R}^d \to [0, 1]$
    - Black box classifier that takes a data point x with d features
    - Returns the probability that x belongs to a certain class
  - $\phi \in \mathbb{R}^d$
    - Explanation in terms of feature importances for the prediction $f(x)$
    - i.e. coefficients ϕ are treated as the feature contributions to the black box prediction
  - ϕ captures the coefficients of a linear model
  - Let $\mathcal{Z}$ be a set of N randomly sampled instances (perturbations) around x
  - Proximity between x and any $z \in \mathcal{Z}$ is given by $\pi_x(z) \in \mathbb{R}$
  - Vector of these distances over N perturbations in $\mathcal{Z}$ as $\Pi_x(\mathcal{Z}) \in \mathbb{R}^N$
  - Let $Y \in [0, 1]$ be the vector of the black box predictions $f(z)$ corresponding to each of the N instances in $\mathcal{Z}$

Method

- **Lime & KernelSHAP**
  - Model-agnostic local explanation approaches that explain predictions of a classifier f by learning a linear model locally ϕ around each prediction (i.e. $y \sim \phi^T z$)
  - Objective function
    - Explanation that approximates the behavior of the black box accurately in the vicinity (neighborhood) of x

$$\arg\min_{\phi} \sum_{z \in \mathcal{Z}} [f(z) - \phi^T z]^2 \pi_x(z)$$

  - Closed form solution from objective function

$$\hat{\phi} = (\mathcal{Z}^T \mathbf{diag}(\Pi_x(\mathcal{Z}))\mathcal{Z} + \mathbb{I})^{-1}(\mathcal{Z}^T \mathbf{diag}(\Pi_x(\mathcal{Z}))Y)$$

  - LIME
    - Chosen $\pi_x(z)$ heuristically: Cosine or l_2 distance
  - KernelSHAP
    - Game theoretic principles to compute $\pi_x(z)$, guaranteeing that explanations satisfy certain properties

Method

- **Constructing Bayesian Local Explanations**
  - Model the black box prediction of each perturbation z as a linear combination of the corresponding feature values $\phi^T z$ plus an error term ϵ
  - Weights of linear combination ϕ capture feature importances & constitute explanation
  - ϵ captures the error that arises due to mismatch between explanation ϕ & local decision surface of the black box model f

$$y|z, \phi, \epsilon \sim \phi^T z + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})$$

$$\phi|\sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2)$$

Method

- **Constructing Bayesian Local Explanations**
  - Error term is modeled as a Gaussian whose variance relies on proximity function $\pi_x(z)$
  - Proximity function $\pi_x(z)$
    - Perturbations closer to the data point x are modeled accurately
    - Allows more room for error in case of perturbations that are farther away
    - Cosine or l_2 distance or game theoretic principles similar to LIME & KernelSHAP

  - Distributions on error ϵ and feature importance ϕ both consider the parameter σ^2
  - The prior on the feature importances considers σ^2 has an intuitive interpretation
    - If we have prior knowledge that the error of the explanation is small
    - Expect to be more confident about the feature importances

$$y|z, \phi, \epsilon \sim \phi^T z + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})$$

$$\phi|\sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2)$$

Method

- **Constructing Bayesian Local Explanations**
  - The weighted least squares formulation of LIME and KernelSHAP

$$\arg\min_{\phi} \sum_{z \in \mathcal{Z}} [f(z) - \phi^T z]^2 \pi_x(z)$$

  - Corresponds to the Bayesian version of that of LIME and KernelSHAP with additional terms to model uncertainty

$$\phi | \sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2)$$

  - Feature importance uncertainty
    - The uncertainty associated with the feature importances ϕ
  - Error uncertainty
    - The uncertainty associated with the error term ϵ which captures how well our explanation ϕ models the local decision surface of the underlying black box.

Method

- **Constructing Bayesian Local Explanations**
  - Inference process involves estimating the values of two key parameters: ϕ and σ^2
    - Compute the local explanation as well as the uncertainties associated with feature importances ϕ and the error term ϵ
  - Posterior distributions on ϕ and σ^2 are normal and scaled $\text{Inv-}\chi^2$, respectively, due to the corresponding conjugate priors

$$\sigma^2|\mathcal{Z}, Y \sim \text{Scaled-Inv-}\chi^2\left(n_0 + N, \frac{n_0\sigma_0^2 + Ns^2}{n_0 + N}\right)$$

$$\phi|\sigma^2, \mathcal{Z}, Y \sim \text{Normal}(\hat{\phi}, V_\phi \sigma^2)$$

$$\hat{\phi} = V_\phi(\mathcal{Z}^T \text{diag}(\Pi_x(\mathcal{Z}))Y)$$

$$V_\phi = \left(\mathcal{Z}^T \text{diag}(\Pi_x(\mathcal{Z}))\mathcal{Z} + \mathbb{I}\right)^{-1}$$

$$s^2 = \frac{1}{N}\left[(Y - \mathcal{Z}\hat{\phi})^T \text{diag}(\Pi_x(\mathcal{Z}))(Y - \mathcal{Z}\hat{\phi}) + \hat{\phi}^T\hat{\phi}\right]$$

Method

- **Constructing Bayesian Local Explanations**
  - Estimate of the posterior mean feature importances of LIME and KernelSHAP

$$\hat{\phi} = (\mathcal{Z}^T \mathrm{diag}(\Pi_x(\mathcal{Z}))\mathcal{Z} + \mathbb{I})^{-1}(\mathcal{Z}^T \mathrm{diag}(\Pi_x(\mathcal{Z}))Y)$$

  - Our Estimate of the posterior mean feature importances

$$V_\phi = (\mathcal{Z}^T \mathrm{diag}(\Pi_x(\mathcal{Z}))\mathcal{Z} + \mathbb{I})^{-1}$$

  - If use the same proximity function $\pi_x(z)$ in our framework as in LIME & KernelSHAP, the posterior mean of the feature importance $\hat{\phi}$ output by our framework will be equivalent to the feature importances output by LIME & KernelSHAP

Method

- **Constructing Bayesian Local Explanations**
  - Feature Importance Uncertainty
    - Compute the posterior mean of local feature importances $\hat{\phi}$ using closed form expression
    - Estimate credible interval (measure of uncertainty) around the mean feature importances by repeatedly sampling from the posterior distribution

Closed form expression

$$s^2 = \frac{1}{N}\left[(Y - \mathcal{Z}\hat{\phi})^T \mathrm{diag}(\Pi_x(\mathcal{Z}))(Y - \mathcal{Z}\hat{\phi}) + \hat{\phi}^T\hat{\phi}\right]$$

Method

- **Constructing Bayesian Local Explanations**
  - Error Uncertainty
    - Error term ϵ can serve as a proxy for explanation quality
    - Captures the mismatch between the constructed explanation and the local decision surface of the underlying black box

Equation s^2

$$s^2 = \frac{1}{N}\left[(Y - \mathcal{Z}\hat{\phi})^T \mathrm{diag}(\Pi_x(\mathcal{Z}))(Y - \mathcal{Z}\hat{\phi}) + \hat{\phi}^T\hat{\phi}\right]$$

Student's t distribution

$$\epsilon|\mathcal{Z}, Y \sim t_{(\mathcal{V}=n_0+N)}(0, \frac{n_0\sigma_0^2 + Ns^2}{n_0 + N})$$

- Evaluate the probability density function of the above posterior at 0, i.e., $P(\epsilon = 0)$
- Substituting the value of s 2 computed using equation s^2 into the Student's t distribution
- Perfectly captures the local decision surface underlying the black box
- Operation in constant time, adding minimal overhead to non-Bayesian LIME & SHAP

Method

- **BayesLIME**
  - Obtain the Bayesian version of LIME by setting the proximity function
  - D Distance metric (e.g. cosine or l_2 distance) , n_0 & σ_0^2 to small values (10^{-6})
  - Prior is uninformative
  - Compute feature importance uncertainty & error uncertainty for LIME's feature importances

Proximity function

$$\pi_x(z) = \exp(-D(x,z)^2/\sigma^2)$$

Method

- **BayesSHAP**
  - Obtain the Bayesian version of KernelSHAP by setting uninformative prior on σ^2 , $\pi_x(z)$
  - SHAP method views the problem of constructing a local linear model as estimating the Shapley values corresponding to each of the features
  - Shapley values represent the contribution of each of the features to the black box prediction
  - The measures of uncertainty output by our method BayesSHAP capture the reliability of the estimated variable contributions

Proximity function

$$\pi_x(z) = \frac{d-1}{(d \text{ choose } |z|)|z|(d-|z|)}$$

$\lvert z \rvert$	The number of the variables in the variable combination represented by the data point z

Method

- **BayesLIME & BayesSHAP**
  - Encourage BayesLIME and BayesSHAP explanations to be sparse
    - Use dimensionality reduction or feature selection techniques as used by LIME and SHAP to obtain the top K features
    - Construct our explanations using the data corresponding to these top K features

Method

- **Estimating the Number of Perturbations**
  - Major drawbacks of approaches such as LIME and KernelSHAP
    - Do not provide any guidance on how to choose the number of perturbations, a key factor in obtaining reliable explanations in an efficient manner
  - Leverage the uncertainty estimates output by our framework to compute perturbations-to-go G
  - Perturbations-to-go G : Estimate of how many more perturbations are required to obtain explanations that satisfy a desired level of certainty
    - Predicts the computational cost of generating an explanation with a desired level of certainty and can help determine whether it is even worthwhile to do so
    - The user specifies the confidence level of the credible interval (denoted as α) and the maximum width of the credible interval W

e.g. "width of 95% credible interval should be less than 0.1" corresponds to α = 0.95 and W = 0.1.

Method

- **Estimating the Number of Perturbations**
  - Estimate G for the local explanation of a data point x
    - Generate S perturbations around x (where S is small and chosen by the user)
    - Fit a local linear model using our method
    - Provides initial estimates of various parameters shown in equations

Equations

$$\hat{\phi} = V_\phi(\mathcal{Z}^T \mathrm{diag}(\Pi_x(\mathcal{Z}))Y)$$

$$V_\phi = \left(\mathcal{Z}^T \mathrm{diag}(\Pi_x(\mathcal{Z}))\mathcal{Z} + \mathbb{I}\right)^{-1}$$

$$s^2 = \frac{1}{N}\left[(Y - \mathcal{Z}\hat{\phi})^T \mathrm{diag}(\Pi_x(\mathcal{Z}))(Y - \mathcal{Z}\hat{\phi}) + \hat{\phi}^T\hat{\phi}\right]$$

Method

• Estimating the Number of Perturbations

- Given S seed perturbations,
- The number of additional perturbations required G to achieve credible interval width W of feature importance for a data point x at user-specified confidence level α can be computed as:

Perturbations-to-go G

$$\left[\frac{W}{\Phi^{-1}(\alpha)}\right]^2 = \mathrm{Var}(\phi_i) = \frac{4s_S^2}{\bar{\pi}_S \times (G+S)} \implies G(W,\alpha,x) = \frac{4s_S^2}{\bar{\pi}_S \times \left[\frac{W}{\Phi^{-1}(\alpha)}\right]^2} - S$$

- Average proximity $\pi_x(z)$ for the S perturbations $\bar{\pi}_S$
- Empirical sum of squared errors between black box & local linear model predictions s_S^2
- The two-tailed inverse normal CDF at confidence level α

Method

- **Focused Sampling of Perturbations**
  - If Perturbations-to-go G is large, Need to reduce this cost
    - Focused sampling which leverages uncertainty estimates to query the black box in a more targeted fashion (instead of querying randomly)
  - Inspired by active learning, focused sampling strategically prioritizes perturbations whose predictions the explanation is most uncertain about
    - Query the black box only for the predictions of the most informative perturbations
    - Learn an accurate explanation with far fewer queries to the black box
  - Determine how uncertain our explanation ϕ is about the black box label for any given instance Query z
    - Compute the posterior predictive distribution for z given as $\hat{y}(z)|\mathcal{Z}, Y \sim t_{(\mathcal{V}=N)}(\hat{\phi}^T z, (z^T V_\phi z + 1)s^2)$
  - The variance of this three parameter student's t distribution

$$\text{var}(\hat{y}(z)) = ((z^T V_\phi z + 1)s^2)(N/(N-2))$$

Part 3.

Method

- **Focused Sampling of Perturbations**
  - The variance of this three parameter student's t distribution

$$\text{var}\left(\hat{y}(z)\right) = \left(\left(z^T V_\phi z + 1\right)s^2\right)\left(N/(N-2)\right)$$

  - Refer to this variance as the predictive variance $\text{var}(\hat{y}(z))$
  - Captures how uncertain our explanation ϕ is about the black box prediction

Algorithm 1 Focused sampling for local explanations

Require: Model f, Data instance x, Number of perturbations N, Number of seed perturbations S, Batch size B, Pool size A, tempurature τ

function FOCUSED SAMPLE
    Initialize $\mathcal{Z}$ with S seed perturbations.
    Fit $\hat{\phi}$ on $\mathcal{Z}$ ▷ Using Eqn (6)
    for $i \leftarrow 1$ to $N - S$ in increments of B **do**
        $\mathcal{Q} \leftarrow$ Generate A candidate perturbations
        Compute $\text{var}(\hat{y}(z))$ on $\mathcal{Q}$ ▷ Using Eqn (11)
        Define $\mathcal{Q}_{\text{dist}}$ as $\propto \exp(\text{var}(\hat{y}(z))/\tau)$
        $\mathcal{Q}_{\text{new}} \leftarrow$ Draw B samples from $\mathcal{Q}_{\text{dist}}$
        $\mathcal{Z} \leftarrow \mathcal{Z} \cup \mathcal{Q}_{\text{new}}$; Fit $\hat{\phi}$ on $\mathcal{Z}$ ▷ Using Eqn (6)
    end for
    return $\hat{\phi}$
end function

Experiment

- **Process**
  - Evaluate proposed framework by first analyzing the quality of our uncertainty estimates
    - i.e., feature importance uncertainty and error uncertainty
  - Assess our estimates of required perturbations G
  - Evaluate the computational efficiency of focused sampling
  - Describe user study with 31 subjects to assess the informativeness of the explanations output by our framework

Experiment

- **Dataset**
  - COMPAS
    - Criminal history, jail and prison time, and demographic attributes of 6172 defendants
  - German Credit
    - Financial and demographic information for 1000 loan applications, each labeled as a “good” or “bad” customer
  - MNIST
    - Handwritten digits dataset
  - Imagenet
    - Select a sample of 100 images of classes French Bulldog, Scuba Diver, Corn, and Broccoli

- **Model**
  - Random forest classifier (sklearn implementation with 100 estimators) as black box models for COMPAS, German Credit
  - 2-layer CNN to predict the digits for MNIST
  - The off-the-shelf VGG16 model as the black box

Experiment

- **Baseline**
  - Generating explanations
    - LIME and KernelSHAP with default settings
  - Images
    - Construct super pixels as described in LIME
    - Use them as features (number of super pixels is fixed to 20 per image)
  - Parameter
    - The desired level of certainty is expressed as the width of the 95% credible interval

Experiment

- **Quality of Uncertainty Estimates**
  - Well calibrated
  - Highly reliable in capturing the uncertainty of the feature importances

95% credible intervals with 100 perturbations include their true values (estimated on 10,000 perturbations)

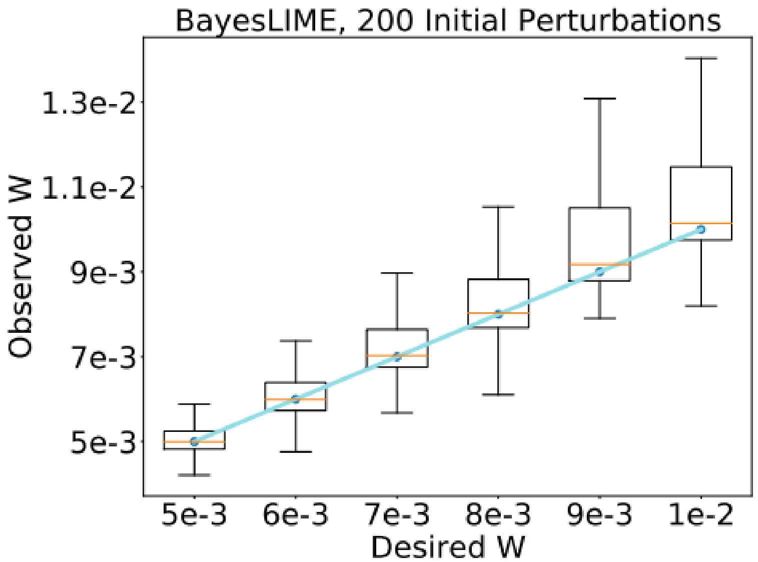
	BayesLIME	BayesSHAP		BayesLIME	BayesSHAP
TABULAR DATASETS			MNIST		
COMPAS	95.5	87.9	Digit 1	95.8	98.4
German Credit	96.9	89.6	Digit 2	95.8	97.4
IMAGENET			Digit 3	95.2	96.3
Corn	94.6	91.8	Digit 4	97.2	90.1
Broccoli	91.4	89.2	Digit 5	95.2	95.6
French Bulldog	94.8	89.9	Digit 6	96.7	96.8
Scuba Diver	92.4	94.6	Digit 7	95.7	95.3

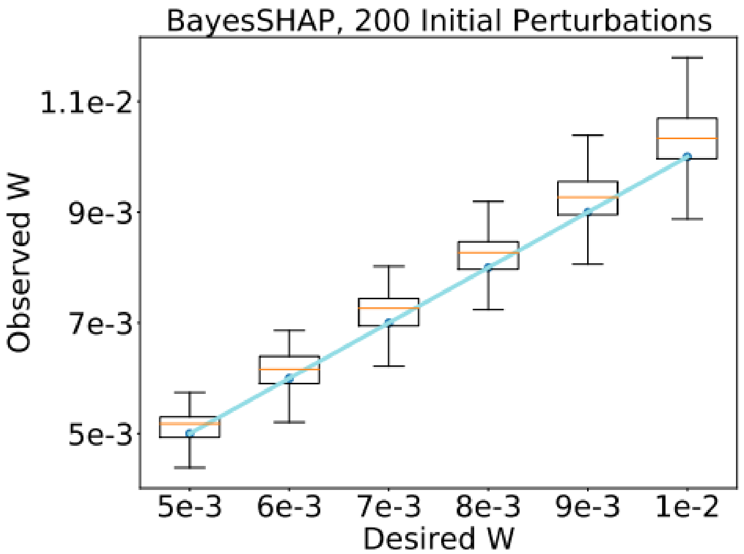
Part 4.

Experiment

- **Correctness of Estimated Number of Perturbations**
  - Leverage these estimates to compute G for 6 different certainty levels
  - Perturbations, where G is computed using the desired credible interval width (x-axis),
  - Compare desired levels to the observed credible interval width (y-axis)
  - Blue line indicates ideal calibration
  - Provides a good approximation of the additional perturbations needed

Perturbations-to-go (G) is averaged over 100 MNIST images of the digit "4"



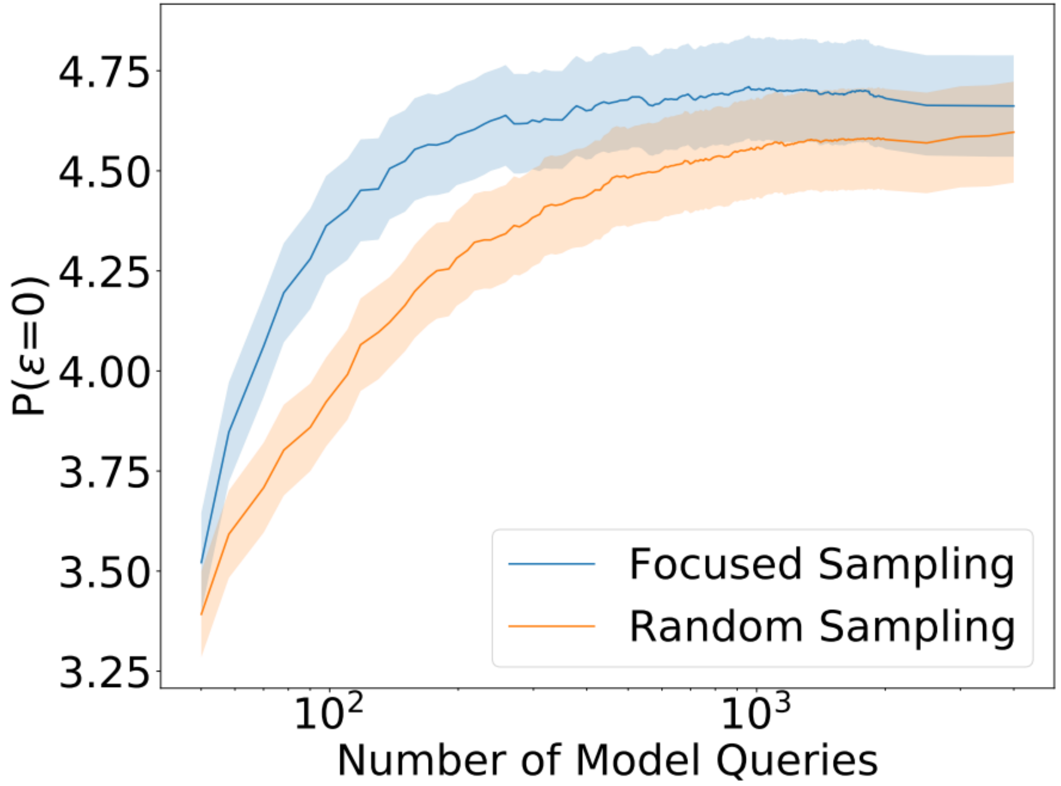


$$\hat{L}(x_i) = \underset{x_j \in N_\epsilon(x_i)}{\text{argmax}} \frac{||\phi_i - \phi_j||_2}{||x_i - x_j||_2}$$

$N_\epsilon(x_i)$ x_i ϕ_i ϕ_j

Experiment

- **Efficiency of Focused Sampling**
  - Results in faster convergence to reliable and high quality explanations
  - Stabilizes within a couple hundred model queries while random sampling takes over 1,000



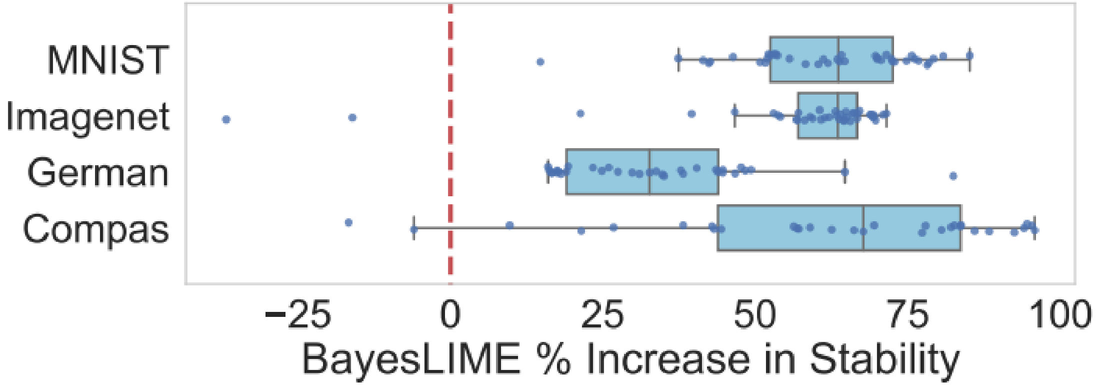
Experiment

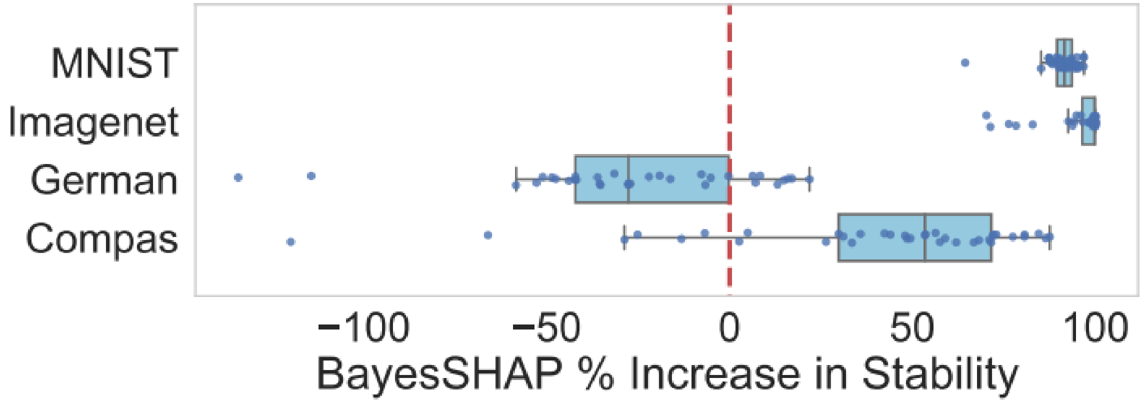
- **Stability of BayesLIME & BayesSHAP**
  - Use the local Lipschitz metric for explanation stability

$$\hat{L}(x_i) = \underset{x_j \in N_\epsilon(x_i)}{\text{argmax}} \frac{||\phi_i - \phi_j||_2}{||x_i - x_j||_2}$$

  - Clear improvement (on average 53%) in stability in all cases except German Credit for BayesSHAP

Assessing the % increase in stability of BayesLIME and BayesSHAP over LIME and SHAP respectively





David Alvarez-Melis et al. On the robustness of interpretability methods. ICML Workshop. 2018.

Conclusion

- **Bayesian framework as solution of previous local explanations**

Problem	**Solution**
Difficult to set hyperparameters	PTG(Perturbations-to-go)
Unclear when you have a good explanation	Credible Intervals
Unstable, re-reruns lead to different explanation s	Credible Intervals
Often naive sampling : Focused sampling	Focused sampling

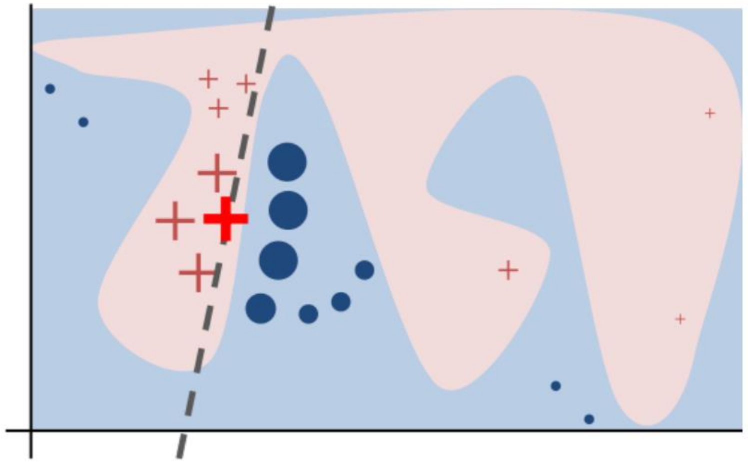
Thank you

2023. 05. 15.

Minseok Yang (msyang0809@khu.ac.kr)
Natural Language Processing Lab
Department of AI, Kyung Hee University

Appendix

- **LIME (Marco Tulio Ribeiro et al., 2016)**
  - Algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with interpretable model



Explanation produced by LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Minimize $\boxed{\mathcal{L}(f, g, \pi_x)}$ while having $\boxed{\Omega(g)}$ be low enough to be interpretable by humans

Formulation

- Explanation family G
- Fidelity function $\mathcal{L}$
- Complexity measure Ω

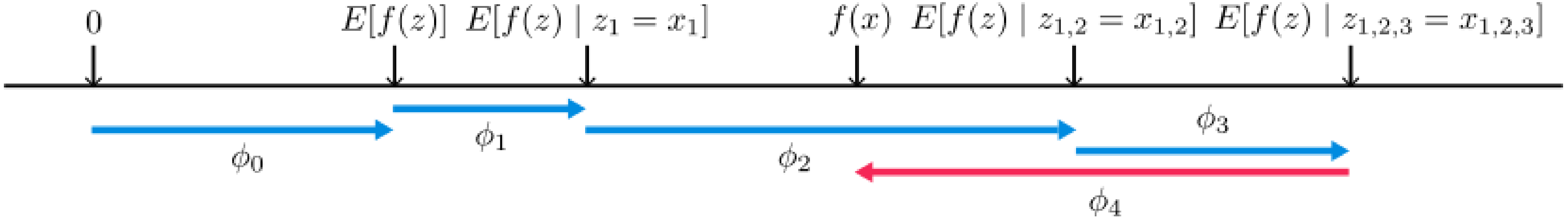
Toy example to present intuition

- Blue/pink background
  - Black-box model's complex decision function (unknown to LIME)
- Red cross
  - Instance being explained
- Dashed line
  - The learned explanation that is locally (but not globally) faithful

Marco Tulio Ribeiro et al. Why Should I Trust You? explaining the predictions of any classifier. KDD. 2016.

Appendix

- **KernelSHAP (Scott M Lundberg et al., 2017)**
  - Game theory results guaranteeing a unique solution apply to th the class of additive feature attribution methods for model explanation



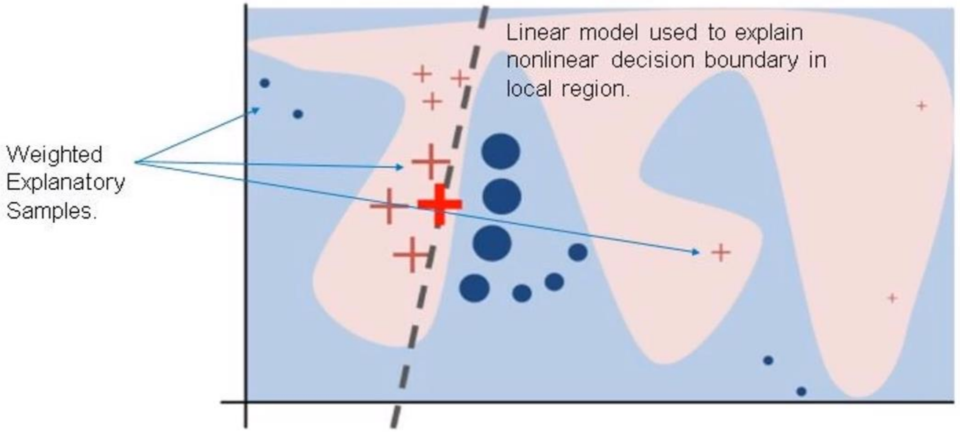
  - SHAP (SHapley Additive exPlanation) values
    - Attribute to feature change in expected model prediction when conditioning on that feature
  - To get from the base value $E[f(z)]$
    - Predict if we did not know any features to the current output
  - When the model is non-linear or the input features are not independent $f(x)$
    - The order in which features are added to the expectation matters
    - SHAP values arise from averaging the ϕ_i values across all possible orderings

Scott M Lundberg et al. A unified approach to interpreting model predictions. NerulIPS. 2017.

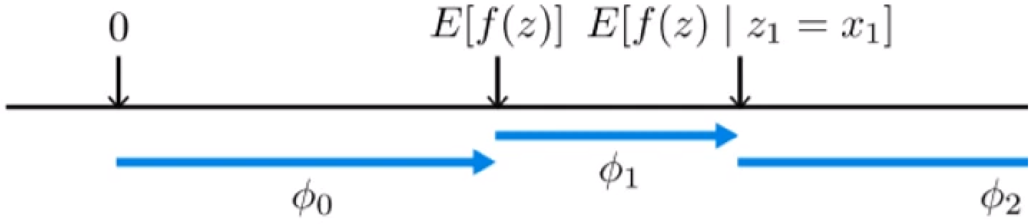
Appendix

- **LIME & KernelSHAP**
  - LIME (Local Interpretable Model-agnostic Explanations)
    - Available on Tabular, Text, Image and even embedding Data
    - Instability because of variation on model explanation
  - SHAP (Shapley Additive Explanations
    - Based on Shapley Values as notion of game theory
    - Kernel SHAP ignore dependence among features

LIME



SHAP

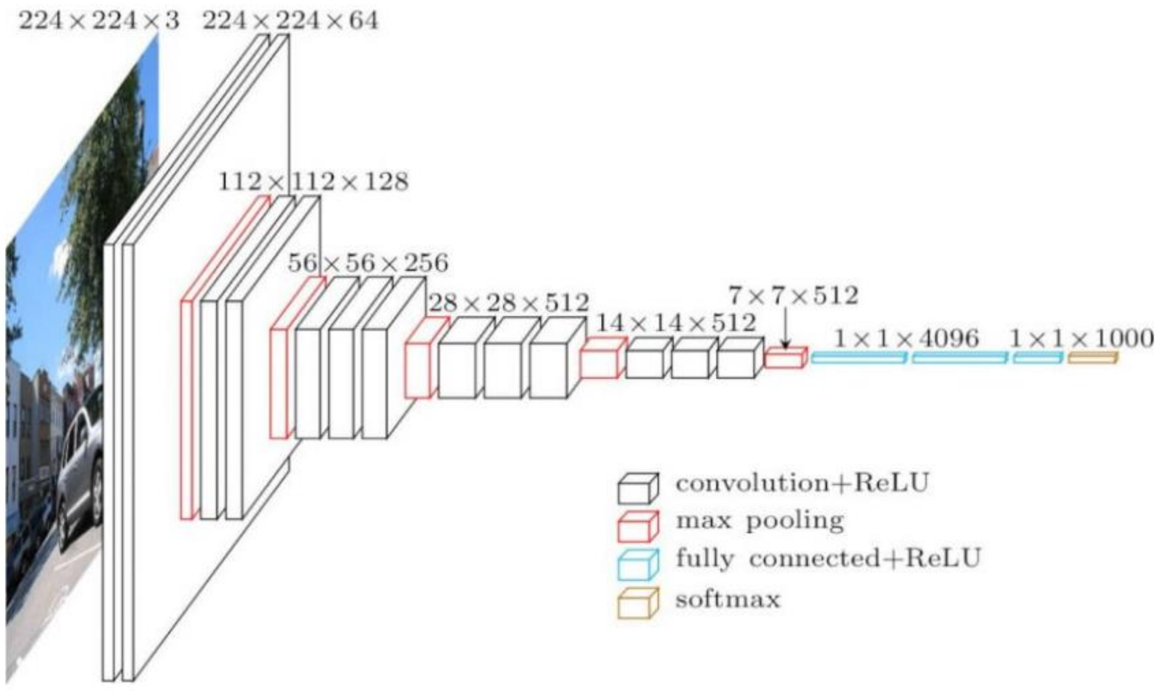


Marco Tulio Ribeiro et al. Why Should I Trust You? explaining the predictions of any classifier. KDD. 2016.
Scott M Lundberg et al. A unified approach to interpreting model predictions. NerulIPS. 2017.

Part 6.

Appendix

- **VGG16 (Karen Simonyan et al., 2015)**
  - Use very small 3 × 3 receptive fields throughout the whole net, which are convolved with the input at every pixel
  - All hidden layers are equipped with the rectification ReLU non-linearity



VGG16 Architecture

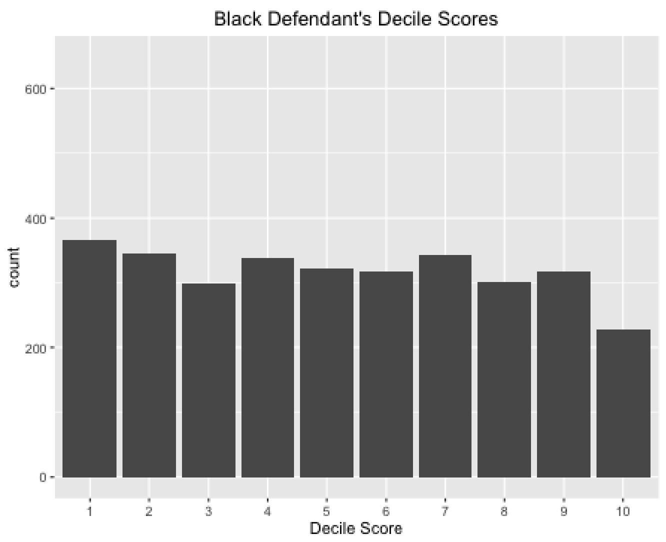
Table 3: **ConvNet performance at a single test scale.**

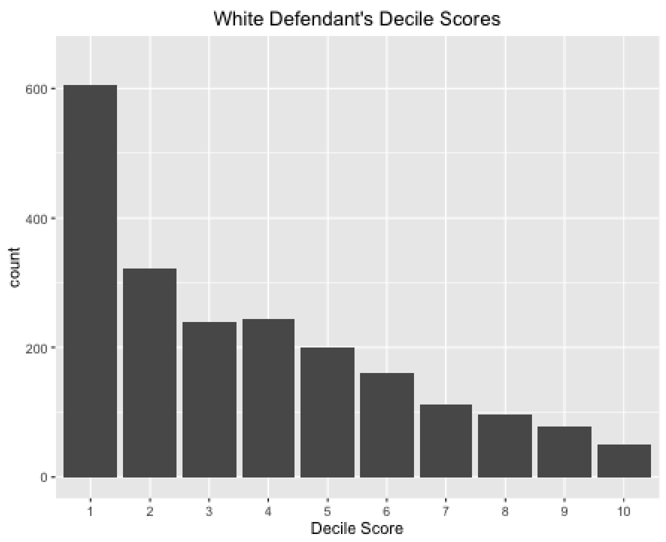
ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. ICLR. 2015.
https://neurohive.io/en/popular-networks/vgg16/

Appendix

- **COMPAS**
  - Northpointe's tool, called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)
    - Discover the underlying accuracy of their recidivism algorithm
    - Test whether the algorithm was biased against certain groups
  - Data production
    - Looked at more than 10,000 criminal defendants in Broward County, Florida
    - Compared their predicted recidivism rates with the rate that actually occurred over a two-year period





Risk of Recidivism

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were

Risk of Violent Recidivism

- To see this analysis result, please visit https://github.com/propublica/compas-analysis

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive. ics.uci.edu/ml.

Appendix

- **German Credit**
  - Classifies people described by a set of attributes as good or bad credit risks
  - Data form
    - Form provided by Prof. Hofmann, contains categorical/symbolic attributes
    - Form provided by Strathclyde University, is used for algorithms that need numerical attributes

Cost matrix

```
..... 1        2
----------------------------
  1   0        1
-----------------------
  2   5        0

(1 = Good,  2 = Bad)
```

Row

- The actual classification and the columns the predicted classification

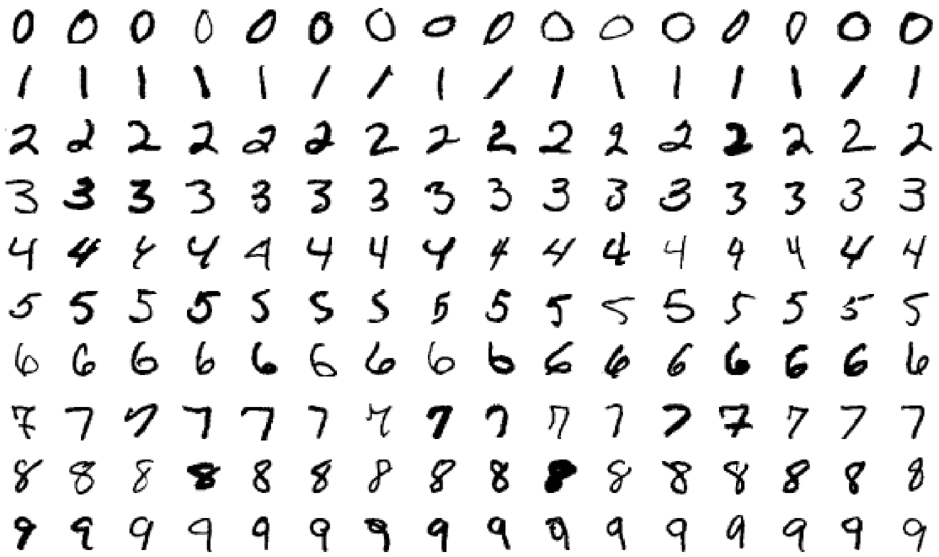
Data Sample

- It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1)

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In ProPublica, 2016.

Appendix

- **MNIST (LeCun et al., 1998)**
  - Task: Image Classification
  - Handwritten number image data
    - Image & label from 0 to 9
    - Image: 28 X 28 pixel
    - Training Data: 60,000 / Test Data: 10,000

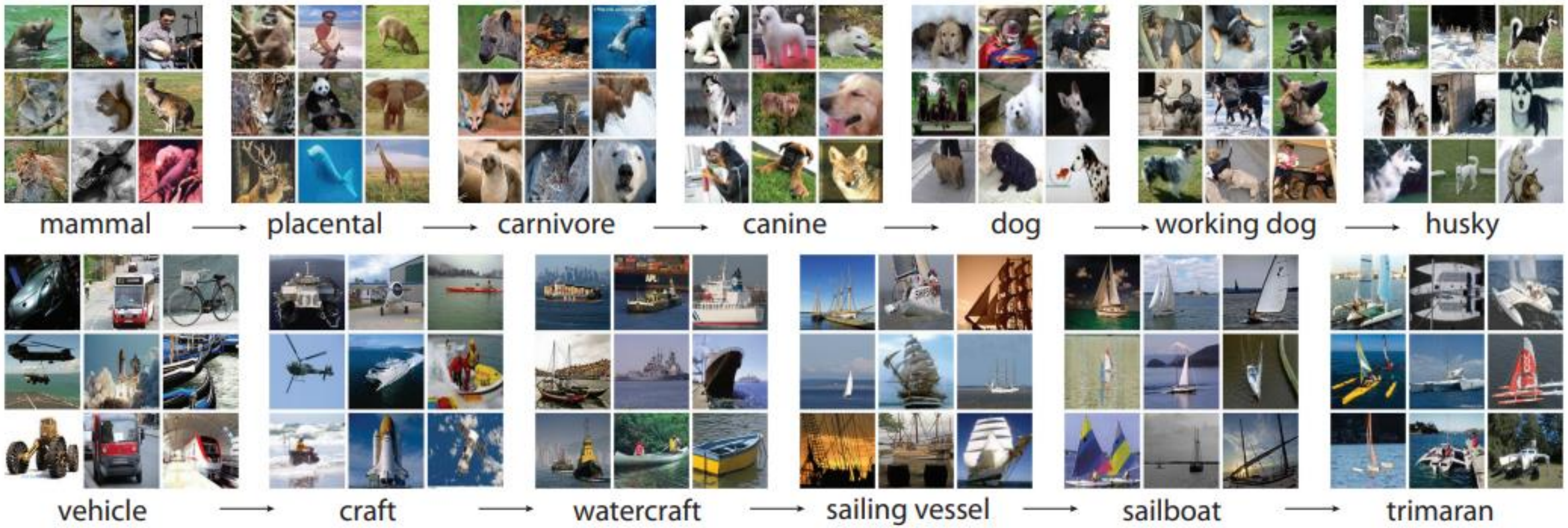


Yann LeCun et al. Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

Part 6.

Appendix

- **ImageNet (J. Deng et al., 2009)**
  - Task: Image Classification
  - Object recognition, image classification and object localization
  - 12 “subtrees”: mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, fruit



J. Deng et al. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.