KERNEL-BASED LIME WITH FEATURE DEPENDENCY SAMPLING

Sheng Shi, Yangzhou Du, Wei Fan

AI Laboratory, Lenovo Research, China

{shisheng2, duyz1 and fanwei2}@Lenovo.com

Abstract

- There are two drawbacks in current existing local explanations. Perturbed samples ignore the intrinsic features correlation. Moreover, most existing methods assume the decision boundary is locally linear.
- We design and develop a novel, high-fidelity local explanation method to address the above challenges. KLFDS: Kernel-based LIME with Feature Dependency Sampling.

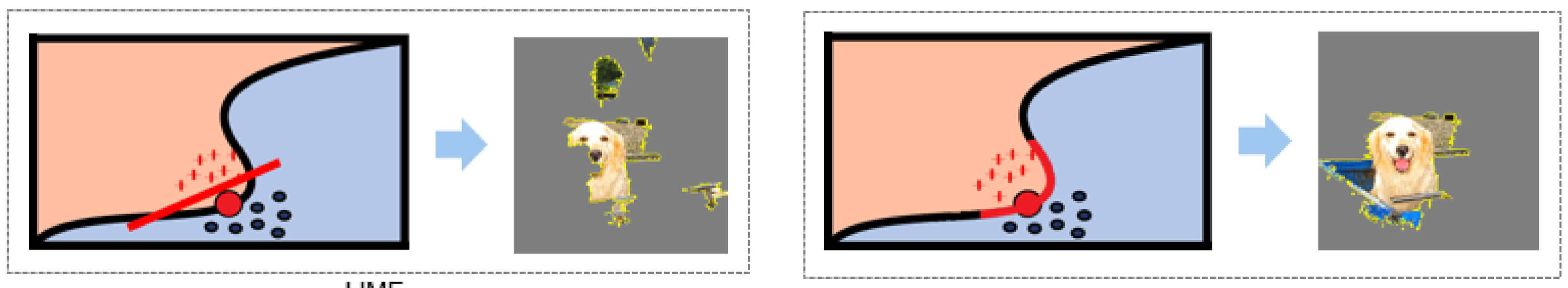
Problem formulation

Problems:

- Perturbed samples ignore the intrinsic correlation between features
  - The visual features of natural objects exhibit a strong correlation in the spatial neighborhood
  - False information contributors lead to poorly fitting of the local explanation model

Problems:

- Most existing methods assume the decision boundary is locally linear.
  - This may produce serious errors as in most complex networks, the local decision boundary is non-linear.



LIME KLFDS

Proposed method

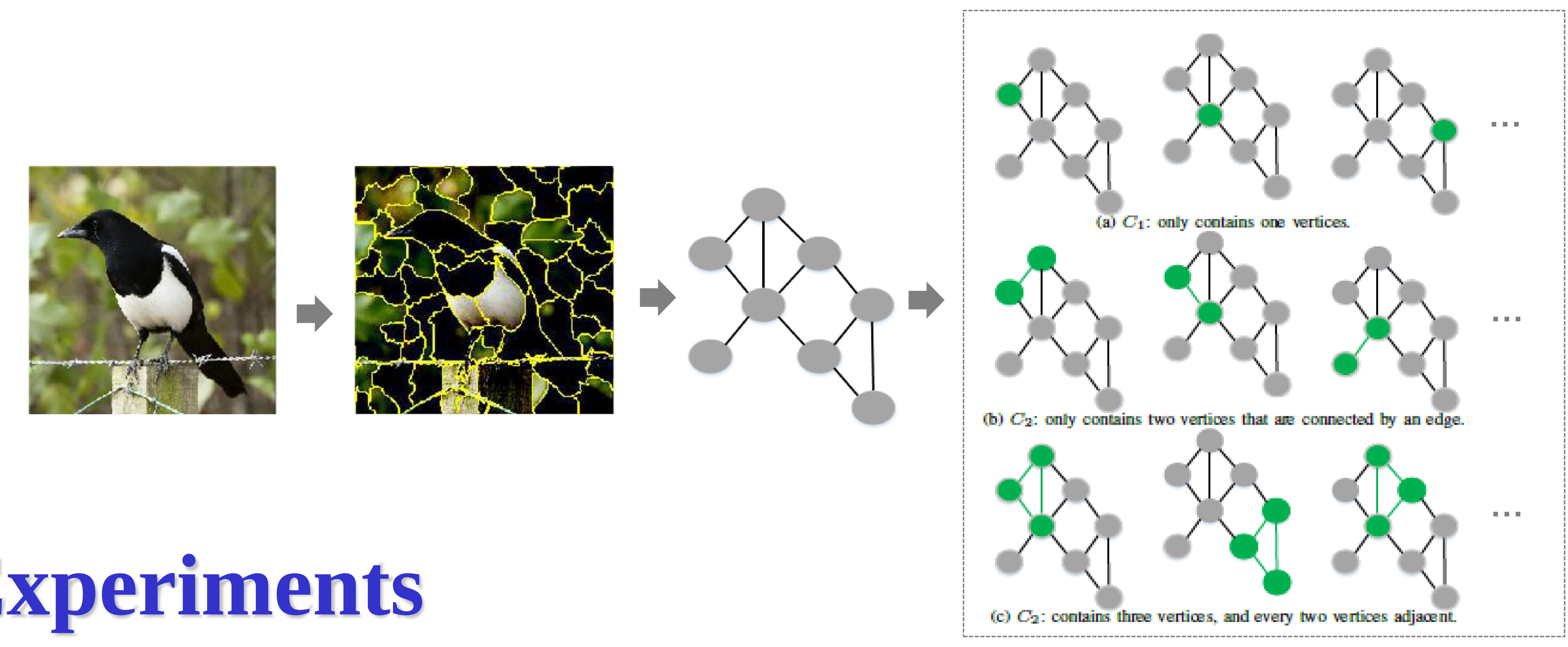
KLFDS: Kernel-based LIME with Feature Dependency Sampling

- Design an unique local sampling process which incorporates the feature clustering method to handle the feature dependency problems.
  - Convert the super-pixel image into an undirected graph
  - Perturbed sampling operation is formalized as clique set construction problem
- Adopt SVR with kernel function to approximate nonlinear boundary.

Algorithm 2 Kernel-based LIME with Feature Dependency Sampling (KLFDS)

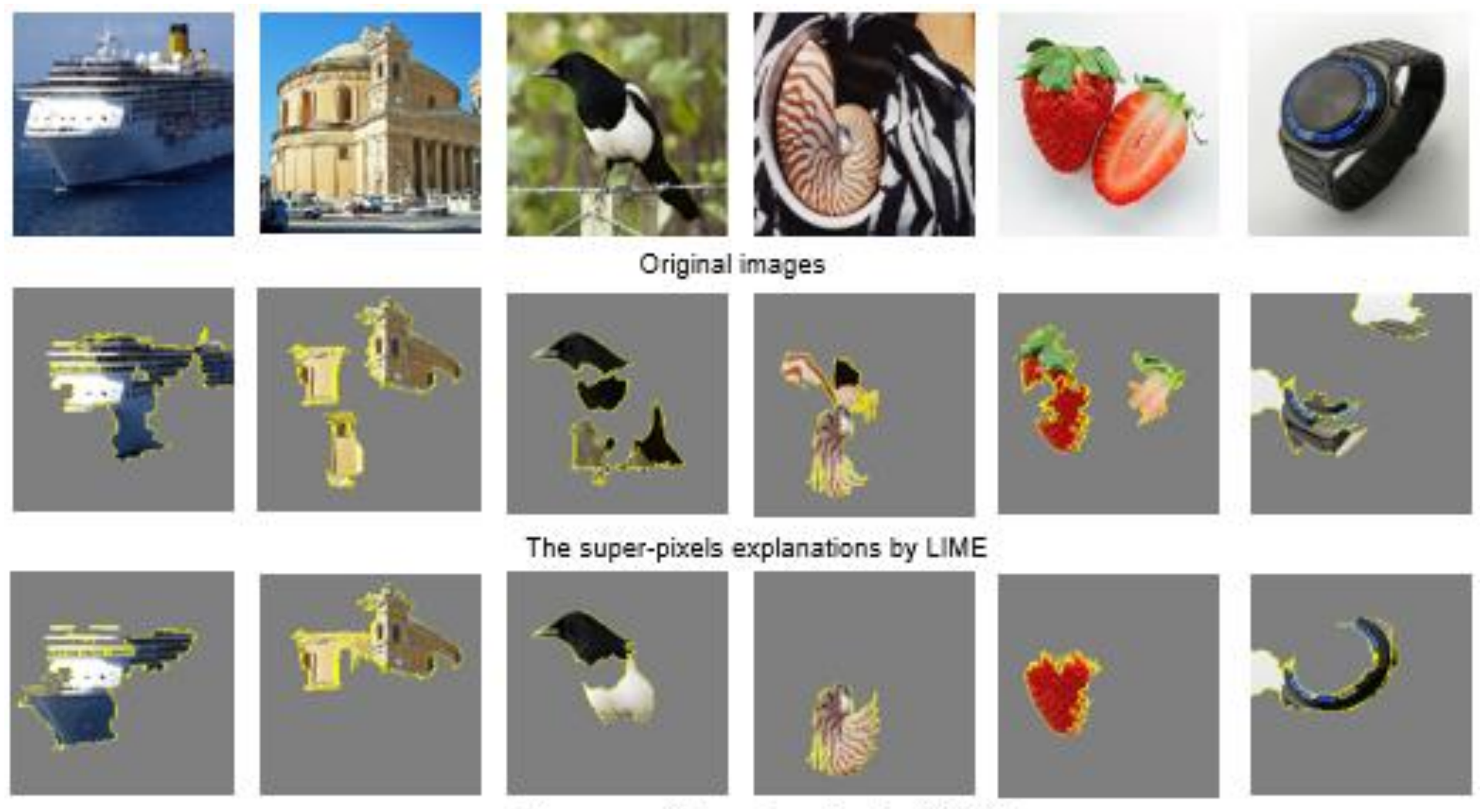
Require: Classifier f, Instance x,

1: get interpretable presentation of x' (e.g. superpixel image for image and bag of word for text)
2: get $f(x')$ by classifier f
3: incorporate the feature clustering method into sampling process to activate a subset of features
4: initial $Z \leftarrow \{\}$
5: **for** $z' \in C$ **do**
6: get z by recovering z'
7: $Z \leftarrow Z \cup (z'_i, f(z_i), \pi_x(z_i))$
8: **end for**
9: use kernel function to project data points into higher dimensional feature space: $g(x, w) = \sum_{i=1}^{N} w_i k(x - x')$.;
10: use the support vector regression to search for a hyperplane
11: **return** feature coefficient



Experiments

- Perform various experiments to explain the Google's pre-trained Inception neural network on Imagenet database.
- Compared with LIME in term of interpretability and fidelity, KLFDS has better performance in explaining classification.



	f(x)	g(x)	Err	R^2
LIME	$P_{castle} = 0.7646$	0.9857	0.2211	0.3219
KLFDS		**0.7633**	**0.0012**	**0.896**
LIME	$P_{yawl} = 0.6076$	0.8129	0.2053	0.4662
KLFDS		**0.6066**	**0.001**	**0.9803**
LIME	$P_{apple} = 0.9943$	1.3028	0.3085	0.5769
KLFDS		**0.9931**	**0.0012**	**0.8118**
LIME	$P_{church} = 0.2886$	0.5133	0.2248	0.4644
KLFDS		**0.288**	**0.0005**	**0.5890**
LIME	$P_{magpie} = 0.9462$	1.2854	0.2655	0.3602
KLFDS		**0.945**	**0.0010**	**0.7955**
LIME	$P_{strawberry} = 0.9797$	1.579	0.5994	0.5299
KLFDS		**0.9784**	**0.0013**	**0.8282**
LIME	$P_{liner} = 0.9669$	1.2422	0.2753	0.6341
KLFDS		**0.9657**	**0.0012**	**0.8414**
LIME	$P_{watch} = 0.9495$	1.0169	0.0674	0.5834
KLFDS		**0.9483**	**0.0013**	**0.9980**
LIME	$P_{nautilus} = 0.9710$	1.3556	0.3846	0.3949
KLFDS		**0.9704**	**0.0006**	**0.8872**

Conclusion

- By simultaneously preserving feature dependency and local non-linearity, KLFDS produces high-interpretability and high-fidelity explanations.