Evaluating Explanations of Convolutional Neural Network Image Classifications

Sumeet S. Shah
Whiting School of Engineering
Johns Hopkins University
Baltimore, USA
sshah97@jh.edu

John W. Sheppard
Gianforte School of Computing
Montana State University
Bozeman, USA
john.sheppard@montana.edu

***Abstract*—In this paper, we seek to automate the evaluation of explanations of image classification decisions made by complex convolutional neural networks (CNN). Explanation frameworks like Local Interpretable Model-agnostic Explanations (LIME) treat complex machine learning models, such as deep neural networks, as black boxes and generate human-interpretable explanations of their decisions using linear proxy models. We propose a pair of experiments to quantitatively evaluate the quality of generated explanations by measuring their sufficiency and salience. To test if a generated explanation contains sufficient information for classification, we test the ability of a trained CNN to classify that explanation properly. We test explanations for salience by training two new CNNs, one using raw image data and the other using explanations as training data, and comparing their classification precision and recall on a common set of test data. We use our new evaluation framework to test our hypothesis that LIME is able to generate explanations that are both sufficient and salient. Our results show that the generated explanations have the potential to be sufficient and salient, provided that the complexity of the explanations is enough to describe the underlying classes.**

***Index Terms*—explainability, sufficiency, salience, LIME**

I. Introduction

In recent years, we have seen the use of machine learning models become more widespread, even as the problems that they are used to address have become increasingly complex. However, in the process of improving models' capabilities and sophistication, we sometimes lose transparency into exactly how they make decisions. Neural networks are prime examples of models that can achieve high performance on complex problems, but lack interpretability, effectively making the trained models black box systems. This is acceptable for applications where the repercussions of mistakes or system exploitation are not too severe; however, for high-risk use cases where undesirable behavior can lead to harm, injury, or death, understanding the system's reasoning process is of paramount importance.

As motivation for our work, we note that just because models such as deep neural networks are not *interpretable* does not mean that they are not *explainable*. While complex models can be difficult for humans to understand, we can use techniques to generate explanations that show which parameters are used for decision making and their relative importance to one another. From this information, we can infer why parameters are being used by considering what we know about the training data and the models' structures. Through these explanations, we can regain some of the transparency of interpretable models without sacrificing performance.

Currently, several methods exist for generating interpretable explanations. However, the quality of these explanations has historically been evaluated by human judges. Being able to evaluate the quality of generated explanations quickly, fairly, and repeatably will be critical as explanation frameworks are developed. In this paper, we present the results of a pair of experiments that demonstrate our ability to evaluate explanations generated by modern explanation frameworks quantitatively in an automated fashion. Specifically, we evaluate the Local Interpretable Model-agnostic Explanations (LIME) [1] framework on its ability to explain image classifications performed by CNNs.

II. Related Work

A. Decision Tree Proxy Models

One of the earlier methods for explaining neural networks was to represent them as decision trees. Initially, the Continuous/discrete Rule Extractor via Decision tree induction (CRED) [2] method was used to translate shallow neural networks. This method was extended by Deep Rule Extractor via Decision tree induction (DeepRED) [3] to handle arbitrarily deep networks. DeepRED uses a number of techniques to prune unnecessary branches from the resulting tree and maximize parsimony. While DeepRED and related decision tree proxy models generate relatively complete explanations, they suffer from high computational complexity and risk generating explanations that are themselves difficult to interpret.

B. Linear Proxy Models

A noteworthy example of a linear proxy method is the Local Interpretable Model-Agnostic Explanations (LIME) framework presented in [1]. LIME is designed to wrap around any black box model. The framework constructs a locally faithful linear model for a given input instance that serves as a proxy for the original global model by performing perturbations on the inputs to the model and observing the results. As this is the explanation framework that we use in our research, we will discuss this method in greater detail in Section IV.

C. Additive Feature Importance

The SHapley Additive exPlanation (SHAP) framework, first described by Lundberg & Lee [4], is relatively new. Presented as a unified, model-agnostic explanation framework, SHAP generates explanations by calculating Shapley values, which are the additive importance that each feature of the input has on the output of the model. Shapley values are a concept that originated in game theory. In that context, they serve as a measure of how important the actions of each player are to the outcome of the game. In the context of machine learning, the players are the input features, and the outcome is the result generated by the model. SHAP's ability to generate intuitive, accurate, and interpretable explanations is demonstrated in [4].

D. Salience Mapping

First conceptualized by Koch & Ullman [5], salience mapping is based on combining visual features that contribute to attentive selection such as color, orientation, etc., into a single map. The salience at a given position in the image is determined primarily by how different that position is from its surroundings in regards to the attentive selectors being considered. The first true implementation of salience mapping was done by Niebur & Koch [6] and was further refined in [7] using color, intensity, orientation, and motion queues as attentive selectors. Techniques such as Randomize Input Sampling for Explanation of black boxes (RISE) [8] have been developed to use salience maps to explain models' behaviors. RISE is similar to LIME in that it treats the model as a black box. As salience maps can be generated independent of a classifier, one can view them as an explanation of how a model should treat a given image. This is the approach taken in [9], where generated salience maps are treated as explanations and used to train new models.

E. Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) is a technique introduced by Bach *et al.* [10] that explains the individual decisions of a model by propagating the prediction from the output backwards to the input. This informs us to what extent the input features affected the final decision. While this technique is geared specifically towards neural network classifiers, Kauffmann *et al.* [11] managed to apply it to other models like clustering and anomaly detection by transforming the models into neural networks and then applying LRP. As it only takes two passes through the model (one forward and one backward) and can be applied to a range of connectionist models, LRP is both flexible and computationally efficient. Unlike proxy model techniques that generate locally faithful models to explain single instances, LRP harnesses the underlying model structure to generate globally faithful explanations.

III. Problem Statement

In spite of the proliferation of explanation generation methods, there is very little standardization for how they are evaluated [12]. Most of the existing explanation frameworks have been evaluated on completeness (compared to the original model), detection of model bias, and subjective human evaluation. For example, both the LIME [1] and SHAP [4] studies evaluated explanations through human evaluation. In contrast, we present a pair of experiments to test the sufficiency and salience of explanations generated using proxy models in an automated fashion.

We say that the explanation of a particular instance is sufficient if it contains enough information to be classified correctly on its own. An explanation of an instance classification decision that is fed back into the original classifier and is itself classified correctly will be considered sufficient. We say that explanations of a class are salient if they contain the most important information for representing the class as learned by the model. To measure salience, we use the generated explanations to train new models and evaluate them. If the explanations used for training contain the most important information from the original classifier, the new models will learn this new information and will be able to classify instances similarly.

We say explanations are complete if they are both sufficient and salient. Explanations that are sufficient and not salient contain enough information to be correctly classified but not enough to represent the feature set that the classifier has learned adequately for a given class. Explanations that are salient and not sufficient contain objective representations of the class but not representations of the class as learned by the model. Lack of either sufficiency or salience is indicative of incomplete explanations.

We limit the scope of this study to CNNs and the LIME framework under the assumption that our experimental design can be adapted to other explanation frameworks and machine learning algorithms with minimal to moderate effort. Given this, we hypothesize that explanations generated by the LIME framework for image classifications made by a CNN will provide sufficient justification for the classifications made, contain salient class information, and will therefore convey a complete representation of the information used to make classification decisions.

IV. Algorithms

A. Inception-v3 CNN

We selected CNN classifiers in our experiments because they represent complex models capable of achieving high performance but lacking native interpretability. We used the Google Inception-v3 framework, which is a 42-layer CNN structure that has been optimized for image classification and object recognition [13]. As the focus of this paper is on our ability to generate good explanations for complex, uninterpretable models, we will not be focusing on the behavior and structure of this model. Instead, we will treat it as a black box, where the input is an image and the output is a set of weights generated by the nodes in the network's softmax output layer, where each node corresponds to a particular class. For training and for classification, every input image is resized to a square 299×299 pixel image without aspect ratio preservation. The training process for each CNN consisted of

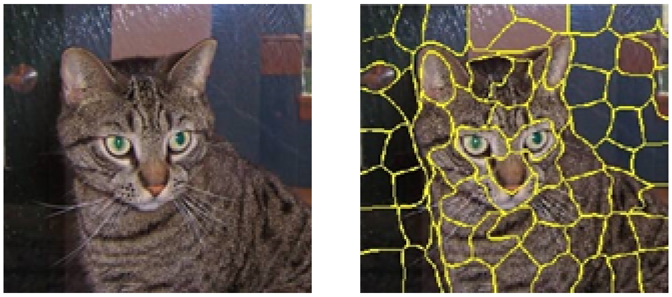

Fig. 1. Image of a cat (left) and Quick Shift segmentation (right)

five 320-step epochs with 32 samples in each gradient-update step.

B. Local Interpretable Model-agnostic Explanations

The LIME framework [1] is an example of a justification-based explainer, meaning that the framework draws connections between the inputs and outputs of the system to generate explanations that provide some degree of justification for the choices made but are not representative of the model's underlying decision process [12]. The framework gathers data on the model by performing a series of perturbations on the inputs and observing the resulting changes in output. This data is used to construct a linear model that serves as a proxy for the original global model in the feature space local to a given input. In the context of the image classification problem, the input perturbations take the form of modified input images with occluded sections. This methodology of finding the most salient input features by systematically occluding portions is similarly applied in salience mapping [14].

How an image is segmented plays an important role in generating explanations, as segmenting the wrong way can cause important features in the image to be divided. Instead of simply dividing the image into a grid, LIME attempts to find meaningful sections in the image using any standard image segmentation technique. Our implementation uses the Quick Shift method, initially introduced in [15]. Quick Shift segments an image by identifying pixel clusters based on spatial and color dimensions. Given an image, the algorithm calculates a forest of pixels where the branches are labeled with a combined spatial and color distance value. Branches with distance values above a predetermined threshold are trimmed, and the remaining sub-trees of contiguous pixels define the segments of the image, which are called "super-pixels." Figure 1 shows how an image of a cat from our data set is segmented into super-pixels.

The LIME framework creates a set of alternative, perturbed images composed of unique permutations of the original image's constituent super-pixels. Each of these perturbed images is fed to the classifier and the outputs are observed. Through this process of systematic image perturbation, LIME is able to derive the importance of each section of the image to the classification decision and construct a proxy model. Because we are using a CNN for classification, the impact of perturbing an image in a particular way can be observed by the changes in the values of the output layer. By observing the impact of all of the perturbed images on the output layer, LIME can assign a weight to each super-pixel for each potential class. That weight, which can be positive or negative, represents the super-pixel's contribution in labeling the instance as that particular class. For our experiments, we configured LIME to create 1,000 perturbed images per classification.

V. Experiments

A. Data

Two data sets were used for our experiments. The first is the Cats & Dogs data set [16]. This data set consists of 25,000 images, exactly half of which are labeled as images of dogs and half of which are labeled as cats. The second data set used is the Flowers data set [17] consisting of 4,242 flower images of five different classes (daisies, dandelions, roses, sunflowers, and tulips) gathered from Flickr, Yandex, and Google Images. The Flower set does not have equal class representation, so the number of images used from each class was reduced to be equal to the least represented class, resulting in 734 images from each class.

B. Design

For each experiment, we performed a 5-fold stratified cross validation. An explanation consists of the set of super-pixels composing an image and their associated weights for each of the possible classes. Within the context of [18], each super-pixel and its weights can be considered to be a "cognitive chunk," i.e., a chunk of salient information. In our experiments, we use the n super-pixels with the highest weights to construct an explanation image. The value of n must be tuned carefully, as explanations with too few super-pixels may be incomplete, while explanations with too many may be noisy and less interpretable.

To test the effect of varying the amount of information in an explanation, we varied the number of super-pixels n. The values of n that we experimented with are 1, 5, 10, 15, 20, and 25. LIME will select up to a maximum of n super-pixels that have a positive weight associated with the predicted class, but will select fewer if there are less than n positively weighted super-pixels for that class.

Figure 2 shows that as we increase the number of super-pixels up to 25, we are allowing approximately half of the original image to be included in the explanation. We did not increase the number of allowed super-pixels further for two reasons. First, we can observe the introduction of several "noise" super-pixels as we increase the allowed count. These noisy super-pixels are determined to be important by the explanation framework but are not relevant to the class being explained. Second, an explanation including over half of the image does not serve as a good explanation if human analysis becomes necessary.

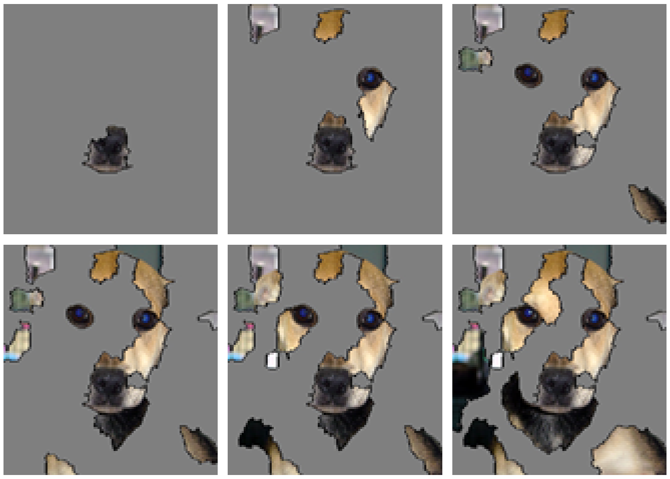

Fig. 2. Explanations with 1, 5, 10, 15, 20, and 25 super-pixels

C. Experiment 1: Sufficient Justification

The first experiment is designed to test if the generated explanations contain enough information to justify the decisions made by the model. First, we trained a CNN on the image data set. We refer to this CNN as the *primary* model. The performance of the trained primary model was evaluated against the data from the test set. Next, for every classification, we generated an explanation for the correct class using LIME. Each LIME explanation was saved as an image where everything except for the features most important for classification were occluded. The primary CNN was then used to classify the set of generated explanations. Explanations containing sufficient justification for the classification of their original image should still be classified correctly. As such, the model's classification precision for each class serves as our measure of how sufficient the generated explanations are.

D. Experiment 2: Salience

The second experiment is meant to test how representative of each class the information contained in the explanations is and if the explanations contain salient representations of the original instances' classes. For this experiment, we trained two new CNNs. The first one, which we refer to as the *secondary* model, was trained using the original test images for which we generated LIME explanations in the sufficiency experiment. The other new CNN, which we refer to as the *explanation* (EX) model, was trained using the LIME-generated explanations themselves. The purpose of the secondary CNN is to serve as a reference point for the EX model. While the training data used in the secondary CNN provides the classifier with a complete set of information per instance (whole images), this is not the case with the EX classifier, where training instances are the explanations with portions of the original image occluded. Note that we did not use the primary model as a reference as it had access to a greater amount of training data.

The performances of both models were evaluated using the primary model's training data as test data. If the EX model is able to perform as well as the secondary model, it would mean that the explanations generated by LIME contain a reasonably salient set of information and indicate that the proxy models constructed by LIME are faithful to the original model. The salience of information transferred to the explanation models via the LIME explanations is observable through their ability to differentiate classes. As precision and recall have equal weight in this test, we use the F-score for each class as a measure of the explanations' salience.

TABLE I
CATS & DOGS PRIMARY MODEL CONFUSION MATRIX

Actual Class	**Predicted Class**	
	Cat	***Dog***
Cat	2500	0
Dog	0	2500

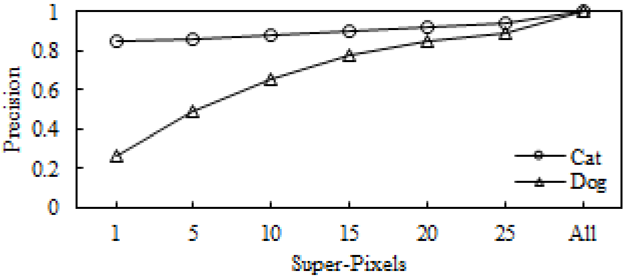

Fig. 3. Cats & Dogs Primary Model Precision on Explanations

VI. SUFFICIENT JUSTIFICATION RESULTS

A. Cats & Dogs

Table I shows the average performance of the Cats & Dogs primary model across the five folds. As shown, the primary model is excellent at correctly classifying both cat and dog images, yielding 100% precision. For most models, perfect performance would be suspicious, or at the very least indicative of an error in the development process. However, given that the Inception-v3 CNN was able to achieve 82.8% precision on the ImageNet data set consisting of 1,000 classes, it is not surprising that it is able to perfectly differentiate between two classes given ample training data.

Figure 3 shows that the ability of the primary classifier to identify the explanations generated from the test data with varying numbers of super-pixels is vastly different between cats and dogs. While the model can identify cats consistently with relatively high precision, it struggles to identify dogs from explanations with fewer super-pixels. However, as we increase the amount of information available in the explanations, the model gets better at correctly classifying dogs, reaching a precision of 90% on explanations containing 25 super-pixels. This upward trend is observed with cats as well, but is far less pronounced. Specifically, cat explanations consisting of a single super-pixel are sufficient to yield 85% precision, and increasing to 25 super-pixels yielded 95% precision.

TABLE II
FLOWERS PRIMARY MODEL CONFUSION MATRIX

Actual Class	**Predicted Class**				
	Daisy	***Dandelion***	***Rose***	***Sunflower***	***Tulip***
Daisy	141	2	0	2	1
Dandelion	41	92	0	6	1
Rose	18	3	73	33	20
Sunflower	59	5	1	80	2
Tulip	25	5	3	25	88

The difference in the number of super-pixels needed for each class to produce a sufficient explanation is not entirely unexpected. Some additional research revealed that, while there are approximately 400 breeds of domestic dogs [19], there are only 40-50 breeds of domestic cats, over 85% of which have arisen within the past century. This means cats have had very few generations to diverge genetically [20]. This suggests that there is much greater in-class variation for dog images, as there are many more species that exhibit more pronounced phenotypic variations. As such, it may be easier for the CNN to isolate features that are distinctly feline than to identify distinctly canine features. Thus there may be features that are shared between both classes that are being interpreted as feline when isolated. The model may be seeing canine traits in the explanations but considers them to be feline when taken out of context of the rest of the image, causing misclassification.

Given that the images of dogs likely have much greater in-class variation than the images of cats, we can assert that the dog class is more complex. It makes sense that a more complex explanation would be needed to explain a more complex class sufficiently. The generated explanations with low super-pixel counts are likely presenting the model with features observed in both canines and felines that the model has learned to be feline based on the lower variability among the cat images. This means that there are features present in the dog image explanations that the model has learned usually belong to cat images. This seems to be a likely explanation for the high sufficiency of cat explanations and low sufficiency of dog explanations at lower super-pixel counts. Increasing the number of super-pixels appears to allow the dog explanations to include less important segments of the image that, in reality, are necessary in addition to the more highly weighted super-pixels to distinguish the dog images.

B. Flowers

Table II shows the performance of the Flowers primary model. Unlike the Cats & Dogs model, which had an abundance of data available to learn just two classes, this model only has 587 instances per class. This primary model does not perform nearly as well, as is to be expected given the limited amount of data available and the difficulty of the classification problem being addressed. The model is able to identify daisies with very high precision, but has relatively low precision for the other 4 classes.

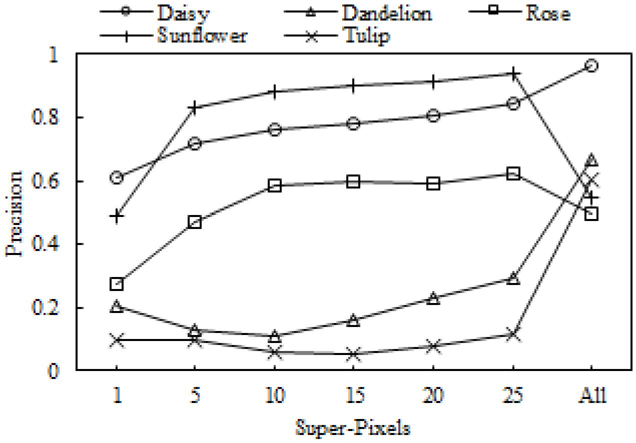

Fig. 4. Flowers Primary Model Precision on Explanations

Figure 4 shows that the ability of the primary classifier to identify the explanations generated from the test data with different numbers of super-pixels varies wildly between the five different classes of flowers. For daisy explanations, we see classification precision climb steadily from 61% to 84% as we increase the number of super-pixels from 1 to 25, while the precision on full daisy images is 97%. Dandelion and Tulip explanations, regardless of the number of super-pixels, are classified with very low precision, being misclassified frequently as daisies, roses, or sunflowers. There is a significant difference in dandelion precision from 29% on 25 super-pixels to 67% on the full images. While there is irregular behavior at low super-pixel counts, we see an approximately linear increase in precision of dandelion explanations from 10 super-pixels to 25 super-pixels. Rose and sunflower explanations follow similar trends to each other. Precision on low super-pixels is relatively poor, but increases significantly with more super-pixels. Interestingly, we see that for both classes, explanations with 25 super-pixels are classified with greater precision than the full images. This is a positive reflection on the performance of the explanation generation framework. Having a greater precision in the explanations suggests that the explanation generation framework is filtering out noise in the original images.

The observed trends in precision suggest that our generated explanations are sufficient for the daisy, rose, and sunflower classes. Furthermore, our generated explanations for roses and sunflowers are more sufficient than the original images. Conversely, even with 25 super-pixels the explanations for dandelions and tulips are mostly insufficient.

C. Sufficient Justification Summary

Based on these results, we have evidence to partially support our hypothesis that the LIME explanations are sufficient. The results of testing the Cats & Dogs image explanations show that they contain sufficient information to be classified correctly; however, the results of testing the Flower image explanations are less clear cut. Three out of the five classes'

TABLE III
CATS & DOGS SECONDARY MODEL CONFUSION MATRIX

Actual Class	Predicted Class	
	Cat	***Dog***
Cat	9984	16
Dog	521	9479

explanations (daisy, rose, sunflower) achieve precision competitive with the classification of the original images with 25 super-pixel explanations, with rose and daisy explanations achieving better precision. However, dandelion and tulip explanations do not approach the precision achieved on the original images. It may be possible to increase the rate of sufficiency of these classes' explanations by raising the restriction on the number of super-pixels allowed in each explanation, but doing so could hamper the human interpretability of the explanations.

It is also possible that raising the super-pixel cap would not help. It may be the case that the Quick Shift segmentation method used to find super-pixels for LIME is overly simplistic for this application and unable to fully represent the features of the dandelion and tulip classes being learned by the CNN. This would explain the large difference in sufficiency observed between rich 25 super-pixel explanations and full images. We would argue that these results show that LIME explanations have the potential to show sufficient justification given that the complexity of the generated explanations is allowed to grow in proportion to the complexity of the class. Even so, a more advanced technique to segment the image, such as the salience mapping techniques described in [5] may be able to better capture the complex features being learned by the model.

VII. SALIENCE RESULTS

A. Cats & Dogs

As a reminder, the secondary models in this experiment were trained using only 5,000 images (2,500 from each class) corresponding to the test images used to evaluate the primary models. The explanations made on these 5,000 images were used to train the EX models. In Table III, we see the performance of the secondary model on the Cats & Dogs images. The model performs very well, achieving a 97% F-score for both cats and dogs. This is competitive with the primary model's 100% F-score, and impressive considering that the secondary model had only $1/5$ of the number of training examples as the primary model. This level of performance indicates that the secondary model has learned to differentiate between cats and dogs very well.

Figure 5 shows the performance of all of the Cats & Dogs EX models, each of which was trained using explanations allowing a different number of super-pixels. We can see that the 1 super-pixel EX model performs poorly for both cats (63% F-score) and dogs (53% F-score). However, once we increase the EX model super-pixel count to 5, the performance increases significantly, rising to 90% for both classes. Further increasing the number of super-pixels used to train the EX models gradually increases the F-score for both classes up to 95% for both classes.

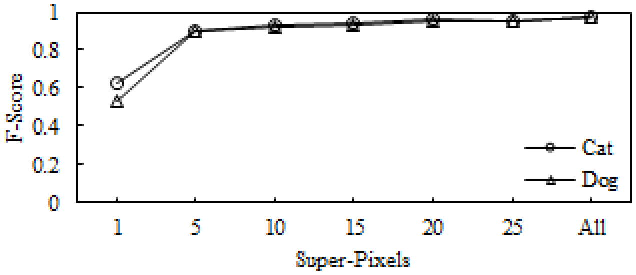

Fig. 5. Cats & Dogs EX Model F-Scores on Original Images

TABLE IV
FLOWERS SECONDARY MODEL CONFUSION MATRIX

Actual Class	Predicted Class				
	Daisy	***Dandelion***	***Rose***	***Sunflower***	***Tulip***
Daisy	563	137	3	11	3
Dandelion	173	369	3	31	6
Rose	49	4	378	89	65
Sunflower	239	13	9	308	14
Tulip	113	11	35	80	245

We can see that the EX model trained with only 5 super-pixels is competitive with the secondary model, indicating that 5 super-pixel explanations contain a salient set of class representation information. This brings up an interesting property of the models' behaviors. Recall that in the previous experiment we found that explanations containing 5 super-pixels were frequently insufficient and misclassified by the primary model. The 5 super-pixel explanations contain salient representations of the underlying classes but insufficient justifications for the primary model's behavior. In simpler terms, the 5 super-pixel explanations are decent representations of the objective cat and dog classes but are incomplete as they do not represent the classes the way that the primary model learned them. They fail to adequately represent the set of features that the primary model used for classification. However, explanations with 20 or 25 super-pixels can be considered complete; they are both sufficient and salient because they can be classified correctly by the primary model and can be used to train accurate EX models.

B. Flowers

In Table IV, we see the performance of the secondary model on the Flowers data set. The model performs well, achieving similar F-scores as the primary model for daisies, sunflowers, and tulips. The secondary model has learned roses slightly better than the primary model, and dandelions slightly worse. As with the Cats & Dogs data, this secondary model had only $1/5$ of the amount of training data as the primary model. However, it should be noted that 20% for Cats & Dogs was still thousands of images, while for this data set 20% amounts to a mere 147 images. These results show that the secondary model has learned to differentiate between flower classes just

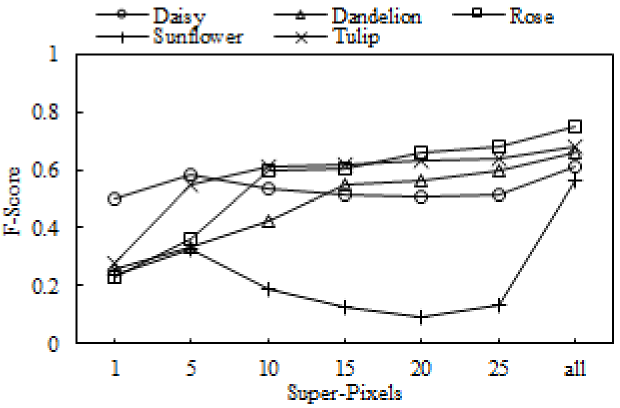

Fig. 6. Flowers EX Model F-Scores on Original Images

as well as the primary model. Once again, the secondary model's performance will serve as a bench mark to compare against the EX models.

Figure 6 shows the performance of all of the Flowers EX models, each of which was trained using explanations allowing a different number of super-pixels. We see a mix of behaviors among the five classes of flowers. For daisies, we can see that the 1 super-pixel EX model has the lowest F-score at 50%, while the secondary model itself has an F-score of 60%, with the other EX models scoring within that range. The EX models are able to classify daisies almost as well as the secondary model, indicating that even 1 super-pixel explanations contain most of the salient information. A manual examination revealed that 1 super-pixel daisy explanations mostly focus on the segments of the images containing the pistil and some of the petals, highlighting the shape and color. The F-score trends of the EX models for dandelions, roses, and tulip, resemble logarithmic curves, similar to those observed for cats and dogs. As was the case with dog explanations, the results of the sufficiency experiment on these three classes also suggest that the explanations with fewer super-pixels are salient representations of the classes but are not representative of feature sets that the primary model uses to classify them, making them incomplete. This is especially pronounced for dandelions and tulips.

The performance of the EX models on sunflower images is unusual. The secondary model has an F-score of 56%. The 1 super-pixel EX model has an F-score of 24%, which improves to 32% with the 5 super-pixel EX model. However, increasing the number of super-pixels in the EX models further causes a significant drop in performance. The EX models have not learned to identify the sunflower class very well, meaning that the explanations used to train them were not salient. This is surprising, as the explanations of sunflower images containing 5 or more super-pixels were mostly sufficient. The EX models are misclassifying sunflower images mostly as daisies. A visual inspection of the generated explanations used to train these models shows that this misclassification is likely caused by similar pistil coloration and petal shape. Since the secondary model performs much better, it appears that the information being transferred to the EX models is not salient, making the explanations largely incomplete.

C. Salience Summary

The results of both experiments on both data sets indicate the LIME framework has potential to generate explanations that are sufficient, salient, and sometimes both. Explanations can be sufficient, justifying the decision made by the original classifier, without being salient (failing to adequately represent the class). Explanations can also be salient and represent the class well, but be insufficient as they do not represent the full set of information for the class as learned by the model. These results provide conditional support for our hypothesis that the LIME framework is capable of generating sufficient and salient explanations, with the condition being that the complexity of the explanations is adequate to capture the complexity of the classes being described. Through these experiments we hope to have set the stage for further experimentation to evaluate explanation generation frameworks quantitatively.

VIII. Future Work

Due to the computationally expensive nature of image processing and explanation generation, we were limited in the scale and number of the experiments that we were able to perform. Altogether, generating the explanations for the Cats & Dogs and Flowers data sets took over 10,000 compute hours. We were able to complete our tests in a reasonable amount of time by harnessing cloud compute resources but were limited in terms of speed and scale by cost. Both of our experiments could be enhanced by running them on a greater number of data sets. Having a wider variety of image classification tasks would help to show the effectiveness of our tests as methods for evaluating explanations. This would serve to highlight the differences in quality of explanations generated as we varied the number of potential classes, in-class variability, between-class variability, and other aspects of the data. We observed our experimental results for the Cats & Dogs data set, which has different in-class variability for each of the two classes, and a medium level of inter-class differences. The Flowers data set was slightly more complex, but we are likely to observe very different results on even more varied and difficult classification problems.

Our observations of the sunflower class from the Flower data set show that 25 super-pixel explanations contain sufficient justification but are not salient. We would theorize that the explanations are not complex enough to properly represent the model's concept of a sunflower. It would be interesting to see how the sufficiency and salience of explanations is affected by the image segmentation technique used in the LIME framework. While Quick Shift was enough for Cats & Dogs as well as most of the Flowers classes, it was not enough for sunflowers. Other segmentation techniques that take more into account than just color and position, such as the salience mapping techniques described in [5], may improve performance.

In the future, we would like to extend our experimental framework to test explanations generated for other learning problems. In this paper, we focused solely on explaining image classification, but there is no reason that we could not run similar experiments on other types of classification or regression problems. We would also like to explore testing other explanation generation methods such as SHAP [4]. While it is also model-agnostic, we specifically avoided the SHAP algorithm due to its high computational complexity. LRP [10] would also be a promising explanation generation technique to test with our evaluation framework, especially considering that, unlike LIME and SHAP, LRP uses the structure of the original model to construct the generated explanations directly. Additional work may be necessary to translate the explanations generated by other frameworks like SHAP and LRP into a format that can be re-entered into the original classifier or used to train a new one. Alternatively, doing so may not be possible, or may not be practical, which would require the modification of our methods to evaluate the sufficiency and salience of explanations.

IX. Conclusion

The results of our experiments provide support for our hypothesis. The explanations generated by the LIME framework have the potential to be sufficient for independent classification and provide salient class representations. Explanations may not always be complete representations of the features used by the classifier to make decisions, but can be, provided that the explanations themselves are allowed to grow in complexity in proportion to the complexity of the underlying classes.

While prior work such as [12] and [18] describe high-level taxonomies for classifying types of explainers and ideas for evaluating them, we have been able to implement concrete evaluation methods borrowing ideas from these higher-level frameworks. We hope that as the field of explainable machine learning continues to expand that our methods will be used to assess the quality of generated explanations and be extended to a broader range of machine learning problems and models.

Acknowledgments

Much of the work reported in this paper was completed as a part of a semester-long project in the 605.746 Advanced Machine Learning class in the Johns Hopkins University's Engineering for Professional's graduate program. We would like to thank students enrolled in this course for their comments, questions, and insights as this work was completed. We also thank the JHU EP program for supporting working professionals to gain research experience through courses such as this one.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[2] M. Sato and H. Tsukimoto, "Rule extraction from neural networks via decision tree induction," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 3, 2001, pp. 1870–1875.

[3] J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRED–rule extraction from deep neural networks," in *International Conference on Discovery Science*, 2016, pp. 457–473.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[5] C. Koch and S. Ullman, "Selecting one among the many: A simple network implementing shifts in selective visual attention." Massachusetts Institute of Technology Artificial Intelligence Lab Memo No. 770, Tech. Rep., 1984.

[6] E. Niebur and C. Koch, "Control of selective visual attention: Modeling the "where" pathway," in *Advances in Neural Information Processing Systems*, 1996, pp. 802–808.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[8] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," *CoRR*, vol. abs/1806.07421, 2018. [Online]. Available: http://arxiv.org/abs/1806.07421

[9] R. Ghaeini, X. Z. Fern, H. Shahbazi, and P. Tadepalli, "Saliency learning: Teaching the model where to pay attention," *Corr*, vol. abs/1902.08649, 2019.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[11] J. Kauffmann, M. Esders, G. Montavon, W. Samek, and K.-R. Müller, "From clustering to cluster explanations via neural networks," *CoRR*, vol. abs/1906.07633, 2019.

[12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *5th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[15] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European Conference on Computer Vision*, 2008, pp. 705–718.

[16] Microsoft Research, "Kaggle cats and dogs dataset," https://www.microsoft.com/en-us/download/details.aspx?id=54765, 2017.

[17] A. Mamaev, "Flowers recognition dataset," https://www.kaggle.com/alxmamaev/flowers-recognition, 2018.

[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *CoRR*, vol. abs/1702.08608, 2017.

[19] B. Tang, Q. Zhou, L. Dong, W. Li, X. Zhang, L. Lan, S. Zhai, J. Xiao, Z. Zhang, Y. Bao *et al.*, "idog: an integrated resource for domestic dogs and wild canids," *Nucleic Acids Research*, vol. 47, no. D1, pp. D793–D800, 2018.

[20] J. Kurushima, M. Lipinski, B. Gandolfi, L. Froenicke, J. Grahn, R. A. Grahn, and L. A. Lyons, "Variation of cats under domestication: Genetic assignment of domestic cats to breeds and worldwide random-bred populations," *Animal Genetics*, vol. 44, no. 3, pp. 311–324, 2013.