DEMOCRATIZING EVALUATION OF DEEP MODEL INTERPRETABILITY THROUGH CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning interpretability tools, such as (Bau et al., 2017; Ribeiro et al., 2016; Smilkov et al., 2017), have been proposed to explain and visualize the ways that deep neural networks make predictions. The success of these methods highly relies on human subjective interpretations, i.e., the ground truth of interpretations, such as feature importance ranking or locations of visual objects, when evaluating the interpretability of the deep models on a specific task. For tasks that the ground truth of interpretations is not available, we propose a novel framework *Consensus* incorporating an ensemble of deep models as the committee for interpretability evaluation. Given any task/dataset, Consensus first obtains the interpretation results using existing tools, e.g., LIME (Ribeiro et al., 2016), for every model in the committee, then aggregates the results from the entire committee and approximates the "ground truth" of interpretations through voting. With such *approximated ground truth*, Consensus evaluates the interpretability of a model through matching its interpretation result and the approximated one, and ranks the matching scores together with committee members, so as to pursue the absolute and relative interpretability evaluation results. We carry out extensive experiments to validate Consensus on various datasets. The results show that Consensus can precisely identify the interpretability for a wide range of models on ubiquitous datasets that the ground truth is not available. Robustness analyses further demonstrate the advantage of the proposed framework to reach the consensus of interpretations through simple voting and evaluate the interpretability of deep models. Through the proposed Consensus framework, the interpretability evaluation has been democratized without the need of ground truth as criterion.

1 INTRODUCTION

Due to the over-parameterization nature (Allen-Zhu et al., 2019), deep neural networks (DNNs) (LeCun et al., 2015) have been widely used to handle machine learning and artificial intelligence tasks, however it is often difficult to understand the prediction results of DNNs despite the very good performance. To interpret the models' behaviors, a number of interpretability tools (Bau et al., 2017; Ribeiro et al., 2016; Smilkov et al., 2017; Sundararajan et al., 2017; Zhang et al., 2019; Ahern et al., 2019) have been proposed to recover or visualize the ways that DNNs make decisions.

Preliminaries. For example, Network Dissection (Bau et al., 2017) uses a large computer vision dataset with a number of visual concepts identified/localized in every image. Given a convolutional neural network (CNN) model for interpretability evaluation, it recovers the visual features used by the model for the classification of every image via intermediate-layer feature maps, then matches the visual features with the labeled visual concepts to estimate the interpretability of the model as the intersection-over-union (IoU) between the activated feature maps and labeled locations of visual objects. Related tools that interpret CNNs through locating importation subregions of visual features in the feature maps have been proposed in (Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018; Wang et al., 2020a).

Apart from investigating the inside of complex deep networks, (Ribeiro et al., 2016; van der Linden et al., 2019; Ahern et al., 2019) proposed to use simple linear or tree-based models to surrogate the predictions made by the DNN model over the dataset through local or global approximations, so as to capture the variation of model outputs with the interpolation of inputs in feature spaces. Then,

with the surrogate model, these methods interpret the DNN model as ways the model uses features for predictions, e.g., ranking of feature importance, and compare the results with the ground truth labeled by human experts to evaluate interpretability. Besides the use of linear interpolation for surrogates, many algorithms, like SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), and PatternNet (Kindermans et al., 2018) have been proposed to estimate the input feature importance as the way to interpret the models, so as to interpret the model predictions by highlighting the importation subregions in the input. In addition to the above methods, (Zhang et al., 2018a; 2019) proposed to learn a graphical model to clarify the process of making decision at a semantic level. Note that obtaining the interpretation of a model is an algorithmic procedure to explain the model (Samek et al., 2017). On the other hand, through the comparing the interpretation results with the human labeled ground truth, the interpretability evaluation aims at estimating the degree to which a human (expert) can consistently predict the model's result (Kim et al., 2016; Doshi-Velez & Kim, 2017).

In summary, the ground truth of interpretation results (usually labeled by human experts) is indispensable to all above methods for interpretability evaluations and comparisons, no matter ways they interpret the models, e.g., visual concepts detecting (Bau et al., 2017), and feature importance ranking for either local (Ribeiro et al., 2016) or global (Ahern et al., 2019) interpretations. While the datasets with visual objects labeled/localized and/or the importance features ranked have offered in some specific areas, the unavailability of ground truths also limits the generalization of these methods to interpret brand new models on the new tasks/datasets ubiquitously. There is thus the need of a method being able to evaluate the interpretability of models on the datasets where the ground truth of interpretation results is not available.

Our contributions. In this paper, we study the problem of evaluating interpretability on the datasets without ground truth of interpretation results. The basic idea of Consensus is to leverage the interpretability of known models as reference to predict the interpretability of new models on new tasks/datasets. Especially, in terms of general purpose perception tasks, we have already obtained a number of reliable models with decent interpretability, such as ResNets, DenseNets and so on. With a new dataset, one could use interpretability tools (Ribeiro et al., 2016; Smilkov et al., 2017) to obtain the interpretation results of these models, then aggregate interpretation results as the reference. Then for any model, one could evaluate interpretability of the model through comparing its interpretation results with the reference.

Specifically, we propose a novel framework named *Consensus* that uses a large number of known models as a committee for interpretability evaluation. Given any task/dataset, Consensus first obtains the interpretation results for every model in the committee using existing interpretation tools, e.g., LIME (Ribeiro et al., 2016), then aggregates the results from the entire committee and reaches the consensus of interpretations through voting. With the *approximated ground truth*, Consensus evaluates the interpretability of each model through matching its interpretation result and the approximated one, and ranks the matching scores of committee members, so as to pursue the absolute and relative interpretability evaluation results without the ground truth of interpretations labeled by human experts. More specifically, we make contributions as follows.

- We study the problem of interpretability evaluation on the datasets without human labeled ground truth of interpretations. To the best of our knowledge, this work is the first to study the problem of evaluating the interpretability of DNNs while the ground truths of interpretation results are not available, by addressing the technical issues of voting and committees.
- We design and implement *Consensus*, a novel interpretability evaluation framework that incorporates a wide range of alternating interpretation tools, such as LIME (Ribeiro et al., 2016), SmoothGrad (Smilkov et al., 2017), to interpret the model as the variation of outputs over the interpolations of inputs (in feature spaces), from massive perspectives (e.g., local or global interpretation, tree-based or linear surrogate and so on), and carries out the interpretability evaluation through the voting based on interpretation results of models in the committee.
- We carry out extensive experiments to validate Consensus on a wide range of models on new ubiquitous tasks/datasets that the ground truth is not available, and exploit the quantifiable metric of model interpretability to report the overall interpretability evaluation results (Section 3). Case studies (Section 4) confirm the effectiveness of Consensus and show the closeness of Consensus based results to the ground truth of interpretations. Robustness analyses (Section 5) further demonstrate the advantage of the committee that the factors including the use of basic interpre-

tation algorithms, the types of networks in the committee, and the size of committee would have few affects on the Consensus based interpretability evaluation results.

2 CONSENSUS: A FRAMEWORK OF INTERPRETABILITY EVALUATION

In this section, we introduce our proposed framework, namely *Consensus*, which incorporates existing interpretations, such as LIME (Ribeiro et al., 2016) and SmoothGrad (Smilkov et al., 2017), to enable DNN interpretability evaluation without the use of human labeled interpretation results as reference/ground truth. Specifically, Consensus generalizes a simple *electoral system*, and consists of three steps: (1) Committee Formation with Deep Models, (2) Committee Voting for Consensus Achievement[1], and (3) Consensus-based Interpretability Evaluation, as follow.

Committee Formation with Deep Models. Given a number of deep neural networks (which are well-known with decent performance on common perception tasks) and a target task with a dataset (potentially without ground truth of interpretation results), Consensus first trains the given neural networks (from scratch) using the dataset. Then, Consensus forms the post-trained networks as a committee of models, and considers the variety of interpretability of models in the committee would establish the references for interpretability comparisons and evaluation. Note that while our research assumes the human labeled ground truth of interpretation results is not available in the given task/dataset for interpretability evaluation, the labels of samples are requested when handling classifications and regression tasks.

Committee Voting for Consensus Achievement. With the committee of trained models and the target task/dataset for interpretation, Consensus first leverages existing interpretability tools, including LIME (Ribeiro et al., 2016) and SmoothGrad (Smilkov et al., 2017) alternatively, to obtain the interpretation results of every model in the committee on every sample in the dataset. Then, Consensus proposes a voting procedure that aggregates the interpretation results from all models in the committee for every sample, and considers the aggregated result as the *quasi-ground-truth* of the interpretation for the sample. For instance, given LIME results on vision tasks which identify the importance of visual features over superpixels in the sample, Consensus normalizes the results of different models into the same scales, and averages the normalized results to obtain the quasi-ground-truth of interpretations for the sample. For SmoothGrad, Consensus follows the similar normalization-averaging procedure to aggregate results of SmoothGrad to obtain the quasi-ground-truth for every sample, which characterizes the feature importance on the pixel-level for vision tasks. To the end, consensus has been achieved through obtaining the collections of quasi-ground-truth for every sample in the target dataset based on committee voting.

Consensus-based Interpretability Evaluation. Given the quasi-ground-truths as the consensus of the whole committee, the proposed algorithm considers the interpretability evaluation of every model in committee as the similarity between interpretation results of the model and the consensus of the whole committee. Specifically, for the interpretation results and consensus based on LIME (visual feature importance in superpixel levels), the proposed algorithm flattens the interpretation results of the model and Consensus for every sample into two vectors, proposes to use cosine of the two vectors as the similarity measure, then Consensus quantifies the interpretability evaluation of the model through the mean of cosine measures over all samples. For the result based on SmoothGrad (visual feature importance in pixel levels), Consensus follow a similar procedure, where the proposed algorithm uses Radial Basis Function (RBF) to tackle the dimensionality issues for the similarity measurement. We rank all models in the committee using their similarities to the Consensus and considers the top/bottom models with good/bad interpretability in the committee.

Note that for any new models for interpretability evaluation, Consensus includes them as members of committee together with a number of known models, and performs above procedures to obtain their interpretability evaluation results (absolute evaluation results) as well as the ranking in the committee (relative evaluation results). In this way, one can clearly position the interpretability and potentials (in terms of performance) of the new models on the new tasks among the known models, even when the ground truth of interpretation results are not available.

[1]We are not intending to connect our work with the multi-agent research though we use the term "consensus achievement".

3 Overall Evaluation and Results

In this section, we use the image classification as the target task for interpretation and interpretability evaluation of deep models. We first introduce the settings of image classification tasks in details, as the setups of our experiments. Then, we present the overall results of Consensus, where we could observe its capacity of evaluating the interpretability of models, with connections to the model performance, while the ground truth of interpretations is not used. Through the comparisons with interpretability evaluation based on LIME and SmoothGrad algorithms using human labeled ground truth, the effectiveness of Consensus has been evaluated.

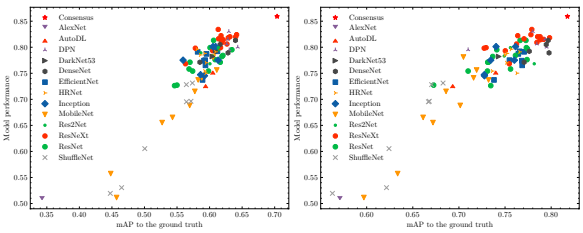
3.1 Evaluation Setups

Here we present the setups of our experiments from following perspectives.

Datasets and Models. For overall evaluation and comparisons, we use two image classification datasets ImageNet (Deng et al., 2009) for ubiquitous visual objects and CUB-200-2011 (Welinder et al., 2010) for birds respectively. Note that ImageNet provides the class label for every image; CUB-200-2011 dataset includes the class label and pixel-level segmentation for the bird in every image, where the pixel annotations of visual objects have been considered as the ground truth of interpretations (Bau et al., 2017). In this way, we evaluate interpretability of models using Consensus, and compare the results with the existing algorithms based on the ground truth.

For the fair comparisons, we use more than 80 models publicly available from PaddlePaddle[2], which have been trained using ImageNet dataset. We also derive models based on CUB-200-2011 dataset through standard training procedures. In our experiments, we include these models in two committees based on ImageNet and CUB-200-2011 respectively for the interpretability evaluation.

Baselines. We consider two interpretation algorithms, LIME (Ribeiro et al., 2016) and SmoothGrad (Smilkov et al., 2017). Specifically, LIME surrogates the interpretation as the assignment of visual feature importance to superpixels (Vedaldi & Soatto, 2008), SmoothGrad outputs the interpretation results as the visual feature importance over pixels. In this way, we can evaluate the flexibility of the proposed Consensus framework over interpretation results from diverse sources (i.e., linear surrogates vs. input gradients) and in multiple granularity (i.e., feature importance in superpixel/pixel-levels). Note that both algorithms use mean Average Precision (mAP) between their interpretation results and the ground truth (i.e., pixel level segmentation, if available) as the measure of interpretability evaluation.



(a) LIME (b) SmoothGrad

Figure 1: Correlation between the Model Performance (Testing Accuracy) and the Ground Truth based Interpretability Evaluation using (a) LIME and (b) SmoothGrad with CUB-200-2011 over 85 models. Pearson correlation coefficients are 0.9269 for LIME (p-value $\ll 0.05$) and 0.9164 for SmoothGrad (p-value $\ll 0.05$). The points "Consensus" here refer to the testing accuracy of the ensemble of networks in the committee by probabilities averaging and voting (in y-axis), as well as the mAP between Consensus results and the ground truth (in x-axis).

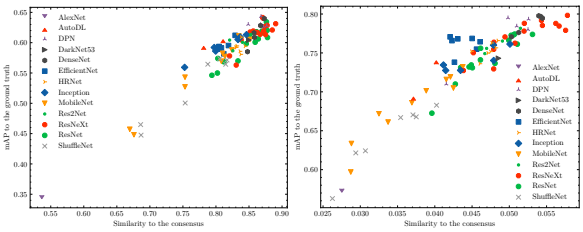
Metrics. Given the similarity measures between the model's interpretation results and consensus of the committee, the proposed algorithm is used to rank the interpretability of every model in the committee. In this way, we compare the ranking list of model interpretability based on Consensus evaluation results with the ranking list based on the ground truth, so as to understand how well the proposed Consensus framework can approximate the evaluation results of model interpretability without the use of ground truth. More specifically, visual comparisons, Pearson correlation coefficients (which characterize the linearity between two variables, e.g., model performance vs. interpretability evaluation) and significant tests have been used as the metrics for overall evaluation and comparisons.

[2] https://github.com/PaddlePaddle/models/blob/release/1.8/PaddleCV/image_classification/README_en.md#supported-models-and-performances

Note that we do not have the ground truth of interpretations for ImageNet. However, it is still possible for us to access the interpretability evaluation through connecting Consensus outputs with the generalization performance of the models. Figure 1 illustrates an example of correlations between interpretability evaluation results based on ground truth and the testing accuracy of 85 models over CUB-200-2011 dataset, where we can see strong, significant, and consistent correlations between model performance and the interpretability.

3.2 Overall Comparisons

In Figure 2, we present the interpretability evaluation results using CUB-200-2011, and the overall comparisons between Consensus and the ground truth based, where we plot the scatter points (i.e., Consensus results in x-axis vs. ground truth based results in y-axis) of the two comparisons using LIME and SmoothGrad respectively. In both comparisons, Consensus performs almost-identically as the one based on ground truth with strong Spearman's correlation (which characterizes the consistency between two ranking lists) with significant tests passed.



(a) Using LIME (b) Using SmoothGrad

Figure 2: Ground-Truth-based v.s. Consensus-based Interpretability Evaluation using (a) LIME and (b) SmoothGrad with CUB-200-2011 Datasets over 85 models (the Committee). Spearman's correlation coefficients are 0.8850 for LIME (p-value $\ll 0.05$) and 0.9063 for SmoothGrad (p-value $\ll 0.05$).

To enable the similar comparisons using ImageNet dataset (where the ground truth of interpretations is not available), we connect the Consensus results with the model performance (testing accuracy) of models, as the model performance and the ground truth based interpretability evaluation results are usually correlated (please see also in Figure 1 and section 3.1). In Figure 2, we present the correlations between Consensus results (without the use of ground truth) and the model performance for both LIME and SmoothGrad using ImageNet and CUB-200-2011 datasets. Specifically, in Figure 3 (a-b) and (d-e), we present the comparisons between model performance (in y-axis) and the Consensus results (in x-axis) using LIME (a,d) and SmoothGrad (b,e) on ImageNet (a,b) and CUB-200-2011 (d,e) respectively.

All correlations here are strong with significant tests passed, though in some local areas of the correlation plots between model performance and interpretability evaluation the trends are not always consistent. It has been observed that some extremely large networks work well with datasets, while they are lack of interpretability (Bau et al., 2017). In this way, we could conclude that, in an overall manner, interpretability evaluation results based on Consensus using both LIME and SmoothGrad over the two datasets are correlated to model performance with significance.

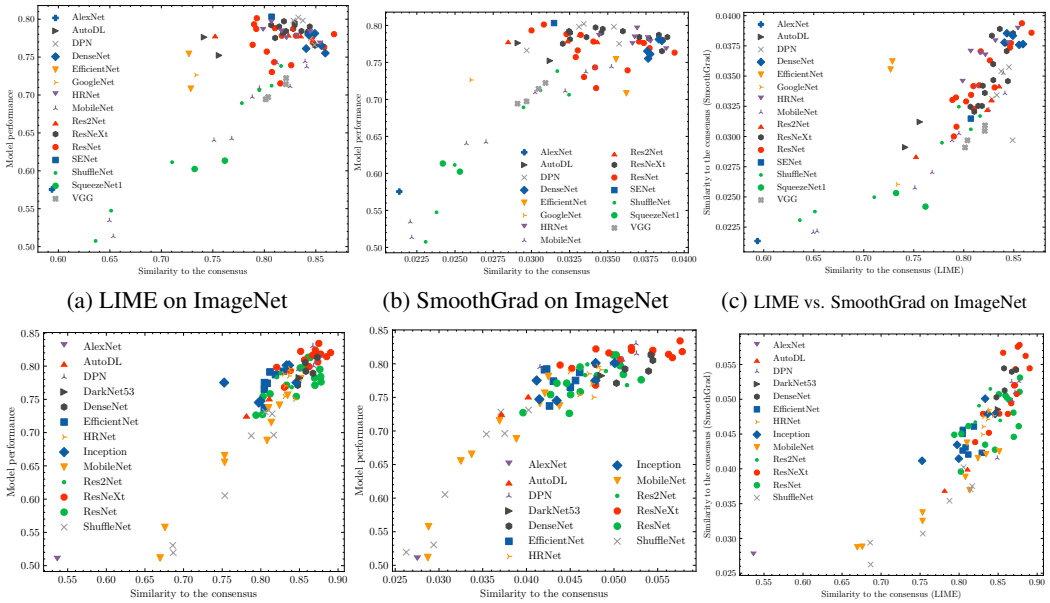
3.3 Comparison Results with Network Dissection (Bau et al., 2017)

In addition to LIME and SmoothGrad, we compare the results of Consensus with ground truth based interpretability evaluation solution – Network Dissection (Bau et al., 2017). With ImageNet dataset, Network Dissection gave a ranking list of models (w.r.t the model interpretability) as follows: ResNet152 > DenseNet161 > VGG16 > GoogLeNet > AlexNet. We report two ranking lists based on Consensus results with detailed numbers for every architecture above, which has been demonstrated in Figure 3 (a, LIME): DenseNet161 (0.849) $\approx$ ResNet152 (0.846) > VGG16 (0.821) > GoogLeNet (0.734) > AlexNet (0.594); and (b, SmoothGrad): DenseNet161 (0.038) $\approx$ ResNet152 (0.037) > VGG16 (0.030) > GoogLeNet (0.026) > AlexNet (0.021). The three ranking lists are almost identical, except the comparisons between DenseNet161 and ResNet152, where in the both lists of Consensus, DenseNet161 is similar to ResNet152 with marginally elevated interpretability, while Network Dissection considers ResNet152 is more interpretable than DenseNet161.

4 Case Studies

In this section, we discuss several technical issues on Consensus for ground-truth-free evaluation of DNN interpretability, using a set of case studies.

Qualification and Effectiveness of Committee-based Voting In our research, Consensus proposes to replace the ground truth with a committee of networks, and it trusts the voting results as



(a) LIME on ImageNet (b) SmoothGrad on ImageNet (c) LIME vs. SmoothGrad on ImageNet

(d) LIME on CUB (e) SmoothGrad on CUB (f) LIME vs. SmoothGrad on CUB

Figure 3: Model performance v.s. similarity to the consensus using LIME (a,d) and SmoothGrad (b,e) over 81 models on ImageNet (a,b) and 85 models on CUB-200-2011 (d,e). The third column shows the similarity to the consensus of SmoothGrad interpretations v.s. similarity to the consensus of LIME interpretations on ImageNet committee (c) and CUB-200-2011 committee (f). Pearson correlation coefficients are (a) 0.8087, (b) 0.7827, (c) 0.8251, (d) 0.9077, (e) 0.8800 and (f) 0.8535.

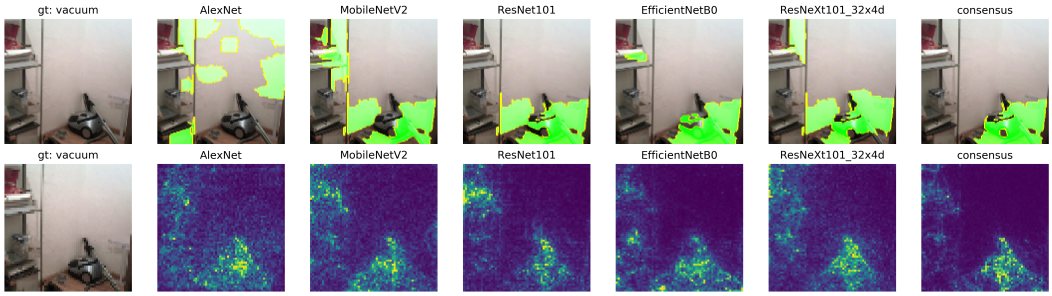


Figure 4: Visual comparisons between Consensus and the interpretation results of CNNs using LIME (in the upper line) and SmoothGrad (in the lower line) based on an image from ImageNet, where the ground truth of interpretation results is not available.

the interpretations of ground truth. Thus, we have to verify (1) whether the (consensus achieved by the) committee would approximate to the ground truth, and (2) whether voting is an effective way to express the interpretation results of the whole committee.

To achieve the goal, we consider the committee as an ensemble of networks, and it classifies data via committee-based voting through averaging the probability outputs of the member networks. In Figure 1, we compare the committee (as an ensemble of networks, entitled with "Consensus") with other architectures for the evaluation of model performance (testing accuracy) and the interpretability evaluation with the ground truth (i.e., mAP between consensus achieved by the committee and the ground truth), using both LIME and SmoothGrad over CUB-200-2011 dataset. The comparison results show that the committee is of the best testing accuracy with significant advantages compared to other models, while the committee is also with the highest mAP between its Consensus interpretations and the ground truth compared to other models. In this way, we can confirm the qualification of the committee, as well as the effectiveness of the committee-based voting.

Closeness of Consensus to the Ground Truth In addition to the mAP measurement (illustrated in Figure 1) between consensus and the ground truth on CUB-200-2011 dataset, we also visualize the

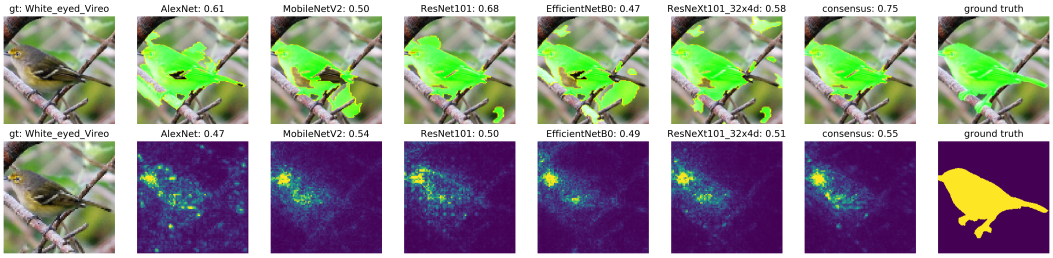


Figure 5: Visual comparisons between Consensus and the interpretation results of CNNs using LIME (in the upper line) and SmoothGrad (in the lower line) based on an image from CUB-200-2011, where the ground truth of interpretation results is available as pixel-wise annotations and the mean Average Precision (mAP) are measured for interpretability evaluation.

examples that compare Consensus achieved by the committee, interpretation results of individual networks using both LIME and SmoothGrad, and (optionally) the ground truth labeled for both ImageNet and CUB-200-2011 in Figures 4 and 5. The comparison shows that Consensus can clearly segment the visual objects related to the classification from the background of images, and it would be closer to the ground truth (if available) than the other networks. Both quantitative results in Figure 1 and the visual comparisons in Figures 4 and 5 validate the closeness of Consensus to the ground truth of interpretations.

5 Robustness Analyses

In this section, we discuss several factors, including the use of basic interpretation algorithms (e.g., LIME and SmoothGrad), the size of committee, and the candidate pool for models in the committee, that would affect the proposed Consensus framework.

Consistency of LIME and SmoothGrad While Consensus adopts LIME and SmoothGrad as the basic interpretation algorithms, the interpretation results from these two algorithms are not exactly the same. Even though the granularity of interpretation results are different, which causes mismatching in mAP estimation with ground truth, the interpretability evaluation results of Consensus based on the two algorithms are generally consistent. The consistency has been confirmed by Figure 3 (c, f), where the overall results of Consensus based on LIME is strongly correlated to SmoothGrad over all models. This shows that the proposed Consensus framework can work well with a wide spectrum of basic interpretation algorithms.

Consistency of Cross-Committee Interpretability Evaluation In real-word applications, the committee-based evaluation makes the results inconsistent in a committee-by-committee manner. In this work, we are interested in whether interpretability evaluation is consistent against the change of committee (e.g., using different sets of models). Given 16 ResNet models as the targets, we form 20 independent committees through combining the 16 models with 10–20 models randomly drawn from the networks presented in Figure 3. In each of these 20 independent committees, we use Consensus to evaluate the interpretability of 16 ResNet models and rank them accordingly. We then estimate the Pearson correlation coefficients between any one of these 20 ranking lists with the list in Figure 3 (a), where the mean Pearson correlation coefficient is 0.96 and the standard deviation is

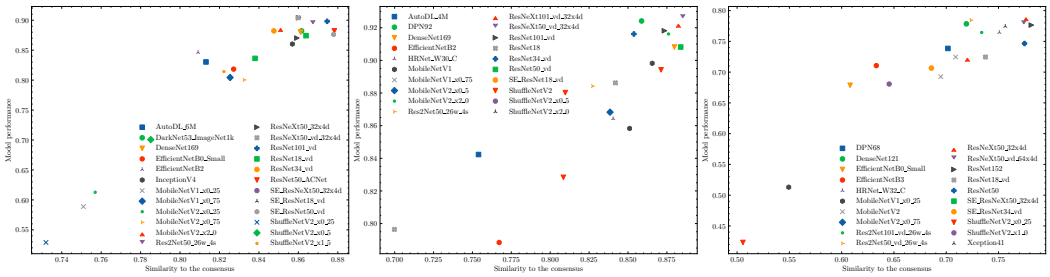


Figure 6: Model performance versus Resules of Consensus using LIME on Stanford Cars 196 (Krause et al., 2013), Oxford Flowers 102 (Nilsback & Zisserman, 2008) and Foods 101 (Bossard et al., 2014). Pearson correlation coefficients are 0.9522, 0.8785 and 0.9134 respectively.

0.04. Thus, we can say the interpretability evaluation based on randomly picked committees would be consistent.

Convergence over Committee Sizes To understand effects of committee sizes to interpretability evaluation, we run Consensus using committees of various sizes formed with networks randomly picked up from the pools. In Figure 7, we plot and compare the performance of Consensus with increasing committee sizes, where we estimate the mAP between the ground truth and the Consensus based on the random committees of different sizes and 20 random trials have been done for every single size independently. It shows that the curve of mAP would quickly converge to the complete the committee, while the Consensus based on a small proportion of committee (e.g., 15 networks) works good enough even compared to the complete committee with 85 networks.

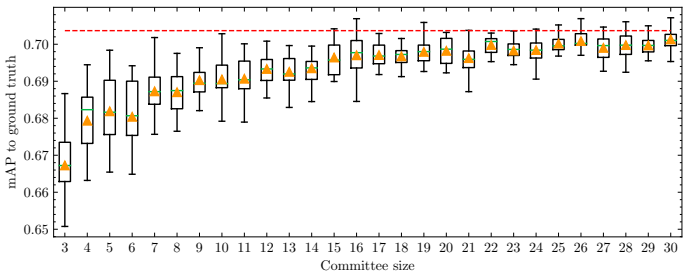


Figure 7: Convergence of mAP between the ground truth and the Consensus results based on committees of increasing sizes, using LIME on CUB-200-2011. The green lines are the mean values and the orange triangles are the median value. The red dashed line is the mAP of Consensus based on the complete committee of the original 85 models.

Applicability with Random Committees over Other Datasets To demonstrate the applicability of Consensus with varying committees over other datasets, we continue our experiments using networks randomly picked-up from the pool with other datasets, including Stanford Cars 196 (Krause et al., 2013), Oxford Flowers 102 (Nilsback & Zisserman, 2008) and Foods 101 (Bossard et al., 2014), in Figure 6, where we consider the connections between the interpretability and model performance, and correlate the results of interpretability evaluation with the testing accuracy (inspired by Figure 1). The results confirm that, when the ground truth of interpretations is not available, our framework is still capable of identifying the interpretability for a wide range of models on ubiquitous datasets/tasks.

6 Conclusion and Future Work

We have proposed a novel framework *Consensus* for evaluating the interpretability of deep models while avoiding the use of ground truth for interpretations. Specifically, Consensus forms a committee of deep models and generalizes an electoral system to reach the consensus of interpretation results using some basic interpretation algorithms via committee-based voting. Then, for every model in the committee, Consensus computes the similarity between its interpretation results and the aggregated consensus, then it ranks the models in the committee accordingly, so as to pursue the absolute score (i.e., similarities to the consensus) and relative results (the rank in the committee) for interpretability evaluation. To validate Consensus, we carry out extensive experimental studies using 85 deep models including ResNets, DenseNets, and so on, on top of 5 datasets, including ImageNet, CUB-200-2011 and so on, in comparison with interpretation algorithms of three categories, including LIME (Ribeiro et al., 2016), SmoothGrad (Smilkov et al., 2017), and Network Dissection (Bau et al., 2017). The results show that (1) Consensus can evaluate the interpretability of models on the datasets even when the ground truth of interpretation results are not available, (2) the consensus of interpretation results aggregated from the committee could well approximate the ground truth of interpretations, (3) the interpretability evaluation results delivered by Consensus correlates to the model performance (testing accuracy) strongly and significantly, (4) the factors including the use of basic interpretation algorithms, the types of networks in the committee, and the size of committee would not affect the results of interpretability evaluation with Consensus.

In terms of methodologies, the proposed Consensus framework considers the segmentation of visual features as interpretations for vision tasks and adopts simple voting mechanism to aggregate results from the committee of models. We believe that the contributions made in this work are complementary with visual objects segmentation and voting. The use of advanced segmentation (Chen et al., 2017a; He et al., 2017) and ensemble learning methods (Dietterich, 2000; Hinton et al., 2015) would further improve the proposed framework while we will include them as part of future work.

REFERENCES

Isaac Ahern, Adam Noack, Luis Guzman-Nateras, Dejing Dou, Boyang Li, and Jun Huan. Normlime: A new feature importance metric for explaining deep neural networks. *arXiv preprint arXiv:1909.04200*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 6541–6549, 2017.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461. Springer, 2014.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. pp. 834–848, 2017a.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4467–4475, 2017b.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1911–1920, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. IEEE, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. URL `http://arxiv.org/abs/1503.02531`.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*, pp. 2280–2288, 2016.

Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. In *International Conference on Learning Representations (ICLR)*, 2018.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 554–561, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pp. 1135–1144, 2016.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, pp. 3145–3153, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. Global aggregations of local explanations for black box models. *FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety - SIGIR 2019 Workshop*, 2019.

Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 705–718. Springer, 2008.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020a.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. 2020b.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *AAAI Conference on Artificial Intelligence*, 2018a.

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6261–6270, 2019.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

A MORE VISUALIZATION RESULTS

We present more visualization results of Consensus, where the samples are from ImageNet and CUB-200-2011.

Figure 8: More visual comparisons between Consensus and the interpretation results of CNNs with LIME on samples from ImageNet, where the ground truth of interpretation results is not available.

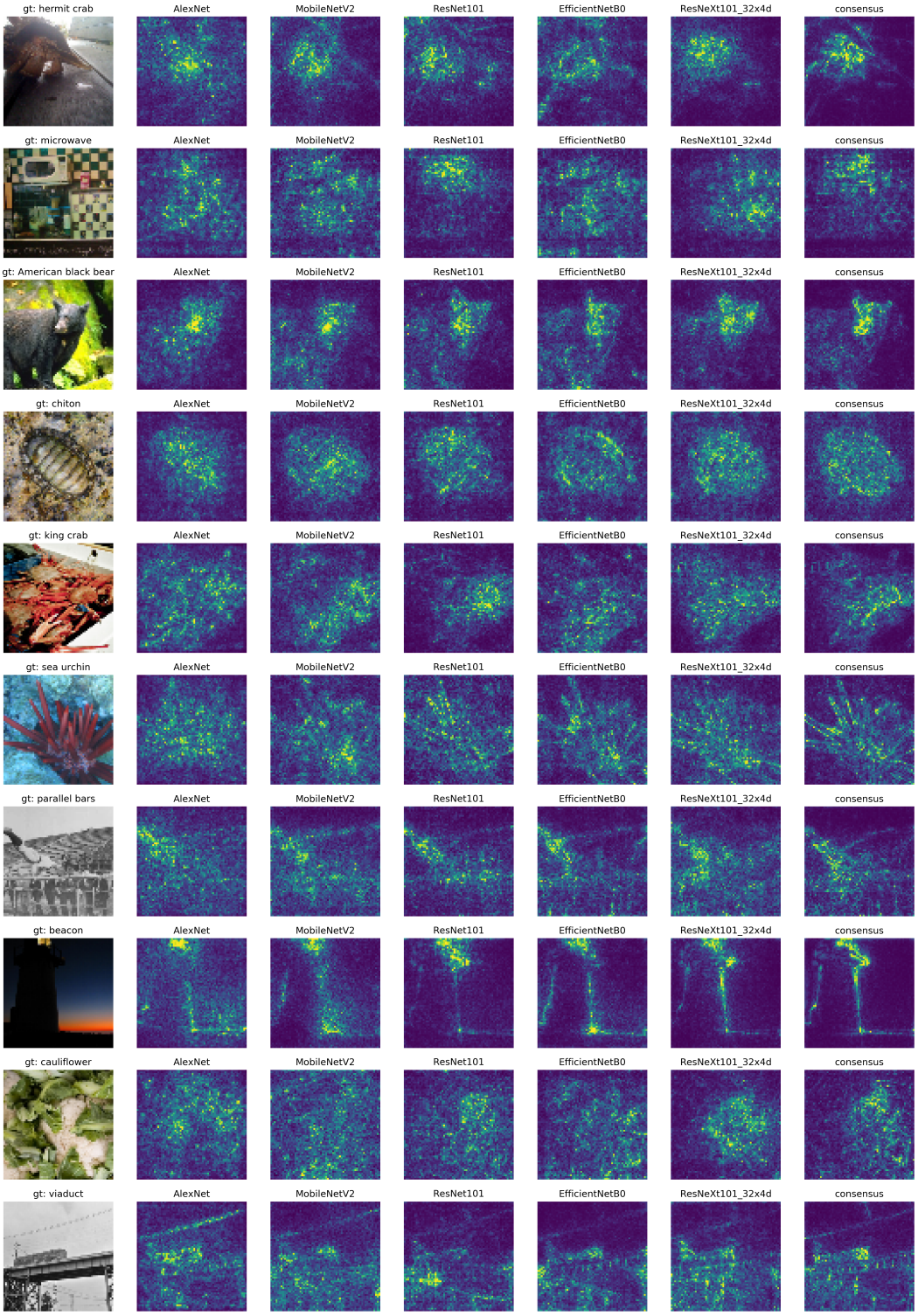


Figure 9: More visual comparisons between Consensus and the interpretation results of CNNs with SmoothGrad on samples from ImageNet, where the ground truth of interpretation results is not available.

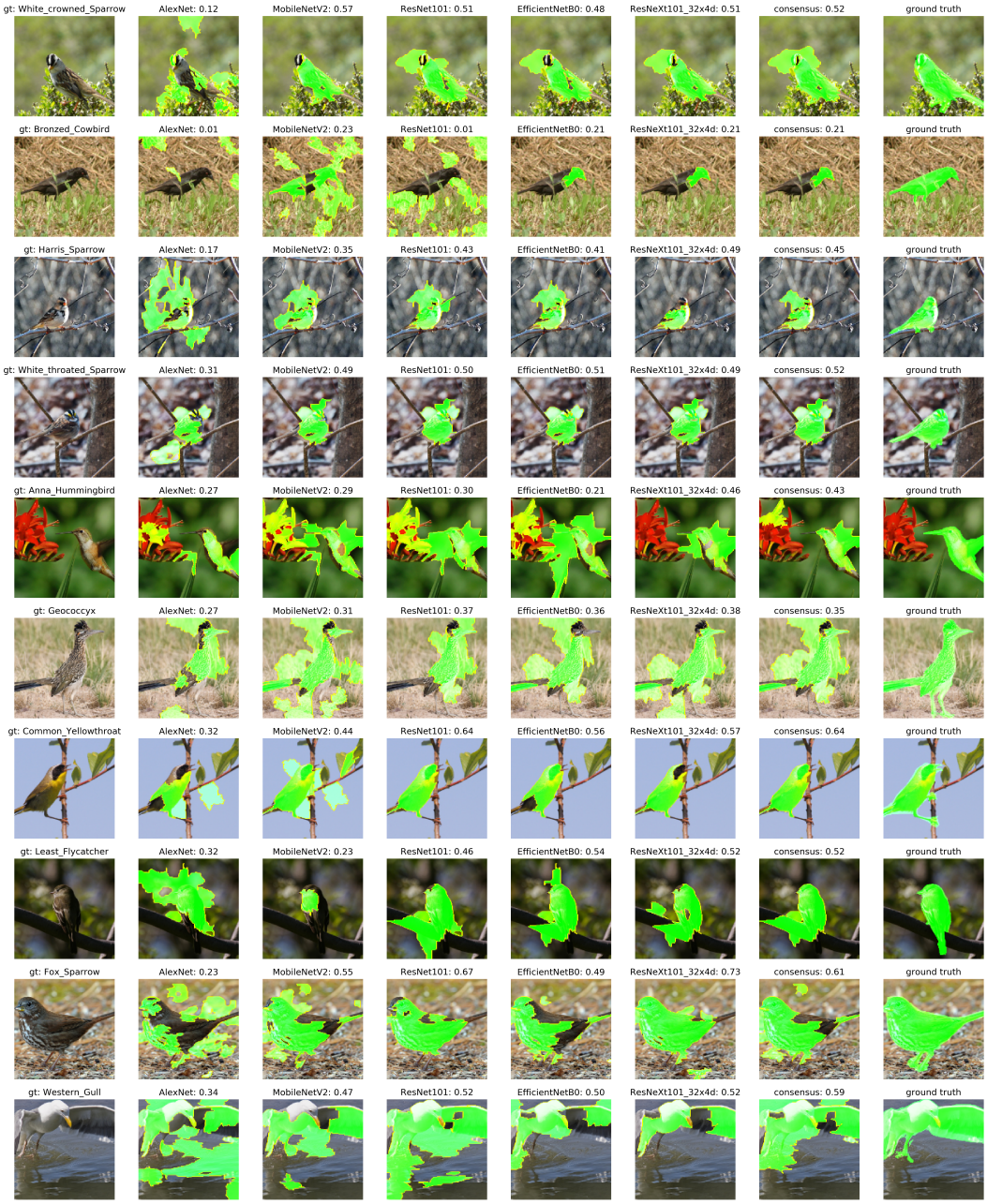


Figure 10: More visual comparisons between Consensus and the interpretation results of CNNs with LIME on samples from CUB-200-2011, where the ground truth of interpretation results is available as pixel-wise annotations and the mAPs are measured for interpretability evaluation.

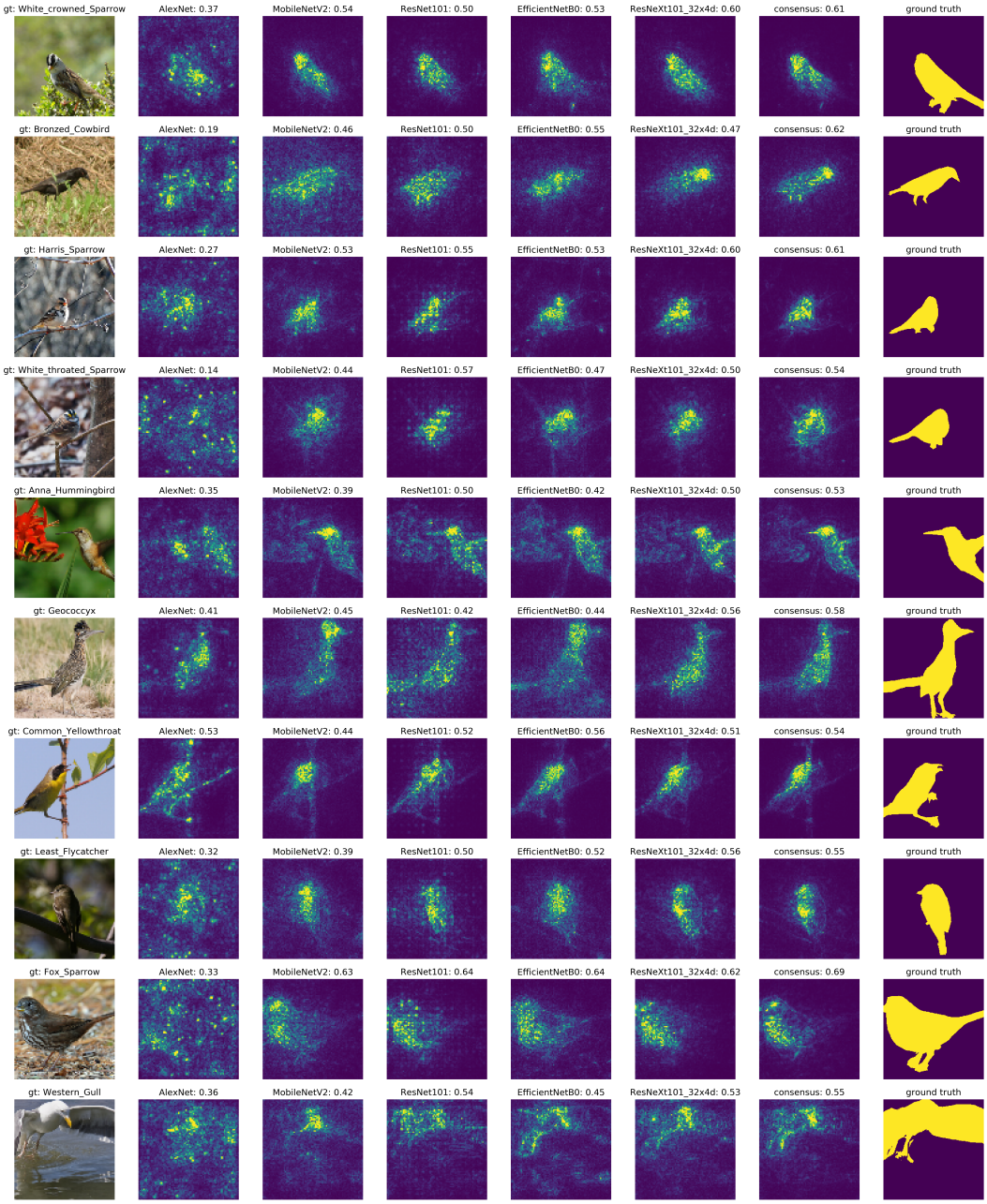


Figure 11: More visual comparisons between Consensus and the interpretation results of CNNs with SmoothGrad on samples from CUB-200-2011, where the ground truth of interpretation results is available as pixel-wise annotations and the mAPs are measured for interpretability evaluation.

B EXPERIMENTAL DETAILS

We present the technique details for the experiments in the main text.

B.1 COMMITTEE FORMATIONS

There are around 100 publicly available deep models trained on ImageNet from PaddlePaddle[3]. We first exclude some very large models that take much more computation resources. Then for the consistency of computing superpixels, we include only the models that take images of size 224×224 as input, resulting 81 models for the committee based on ImageNet. Since there are already a large number of available models, we choose to not include more models by aligning the superpixels in different sizes of images.

As for CUB-200-2011 (Welinder et al., 2010), similarly we first exclude the very large models. Then we follow the standard procedures (Sermanet et al., 2014; Simonyan & Zisserman, 2015) for fine-tuning ImageNet-pretrained models on CUB-200-2011. For simplicity, we use the same training setup for all pre-trained models (learning rate 0.01, batch size 64, SGD optimizer with momentum 0.9, resize to 256 being the short edge, randomly cropping images to the size of 224×224), and obtain 85 models that are well trained. Different hyper-parameters may help to improve the performance of some specific networks, but for the same reason of the large number of available models, we choose to not search for better hyper-parameter settings.

Given the convergence over committee sizes (Figure 7), which shows that a committee of more than 15 models works good enough, thus we randomly choose around 20 models for Stanford Cars 196 (Krause et al., 2013), Oxford Flowers 102 (Nilsback & Zisserman, 2008) and Foods 101 (Bossard et al., 2014), following the same training procedure as CUB-200-2011.

B.2 INTERPRETATION ALGORITHMS

To interpret a model, LIME (Ribeiro et al., 2016) on vision tasks first performs a superpixel segmentation (Vedaldi & Soatto, 2008) for an image, then generates samples by randomly masking some superpixels and computing the outputs through the model, and finally fits the model outputs with the set of superpixels as input by a linear regression model. The linear weights then presents the feature importance in the superpixel level as the interpretation result.

The gradients of model output w.r.t. input can partly identify influential pixels, but due to the saturation of activation functions in the deep networks, the vanilla gradient is usually noisy. SmoothGrad (Smilkov et al., 2017) reduces the visual noise by repeatedly adding small random noises to the input so as to get a list of corresponding gradients, which are averaged for the final interpretation result.

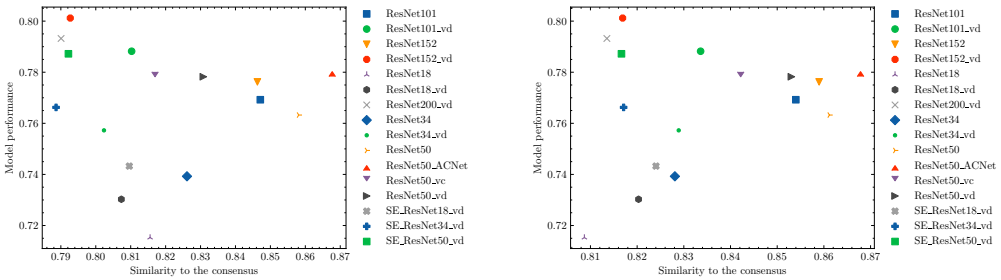
C RESNET FAMILY

We show the zoomed plot of ResNet family in the ImageNet-LIME committee of 81 models in Figure 12 (a). Meanwhile, we also present the results using ResNet family (whose name contains "ResNet" key word) as committee in Figure 12 (b). These two subfigures have no large difference, which confirms the consistency of ranking models in different committees.

D REFERENCES OF NETWORK STRUCTURES

Many structures of deep neural networks have been evaluated in this paper, including AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), ResNeXt (Xie et al., 2017), SEResNet (Hu et al., 2018), ShuffleNet (Zhang et al., 2018b; Ma et al., 2018), MobileNet (Howard et al., 2017; Sandler et al., 2018; Howard et al., 2019), VGG (Simonyan & Zisserman, 2015), GoogleNet (Szegedy et al., 2015), Inception (Szegedy et al., 2015), Xception (Chollet, 2017), DarkNet (Redmon et al., 2016; Redmon & Farhadi, 2018), DenseNet (Huang et al., 2017), DPN (Chen et al., 2017b), SqueezeNet

[3] https://github.com/PaddlePaddle/models/blob/release/1.8/PaddleCV/image_classification/README_en.md#supported-models-and-performances



(a) Based on Complete Committee

(b) Based on ResNet Family

Figure 12: Model performance v.s. similarity to the consensus of LIME on ResNet family. The consensus of (a) is voted by the complete committee on ImageNet (81 models), while the consensus of (b) is voted by ResNet family (16 models).

(Iandola et al., 2016), EfficientNet (Tan & Le, 2019), Res2Net (Gao et al., 2019), HRNet (Wang et al., 2020b), Darts (Liu et al., 2018), AcNet (Ding et al., 2019) and so on.