

# Instance-wise or Class-wise? A Tale of Neighbor Shapley for Concept-based Explanation

Jiahui Li<sup>1</sup>, Kun Kuang<sup>1\*</sup>, Lin Li<sup>1</sup>, Long Chen<sup>2†</sup>, Songyang Zhang<sup>3</sup>, Jian Shao<sup>1</sup>, Jun Xiao<sup>1</sup>

<sup>1</sup>Zhejiang University, <sup>2</sup>Columbia University, <sup>3</sup>University of Rochester  
{jiahuil,kunkuang,mukti}@zju.edu.cn, zjuchenlong@gmail.com  
szhang83@ur.rochester.edu, jshao@zju.edu.cn, junx@cs.zju.edu.cn

## ABSTRACT

Interpreting model knowledge is an essential topic to improve human understanding of deep black-box models. Traditional methods contribute to providing intuitive instance-wise explanations which allocating importance scores for low-level features (e.g., pixels for images). To adapt to the human way of thinking, one strand of recent researches has shifted its spotlight to mining important concepts. However, these concept-based interpretation methods focus on computing the contribution of each discovered concept on the class level and can not precisely give instance-wise explanations. Besides, they consider each concept as an independent unit, and ignore the interactions among concepts. To this end, in this paper, we propose a novel CONcept-based NEighbor Shapley approach (dubbed as CONE-SHAP) to evaluate the importance of each concept by considering its physical and semantic neighbors, and interpret model knowledge with both instance-wise and class-wise explanations. Thanks to this design, the interactions among concepts in the same image are fully considered. Meanwhile, the computational complexity of Shapley Value is reduced from exponential to polynomial. Moreover, for a more comprehensive evaluation, we further propose three criteria to quantify the rationality of the allocated contributions for the concepts, including coherency, complexity, and faithfulness. Extensive experiments and ablations have demonstrated that our CONE-SHAP algorithm outperforms existing concept-based methods and simultaneously provides precise explanations for each instance and class.

## CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Supervised learning by classification**.

## KEYWORDS

Model interpretation; Shapley value; Instance-wise explanation; Class-wise explanation

\* Kun Kuang is the corresponding author.

† This work started when Long Chen at Zhejiang University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475337>

## ACM Reference Format:

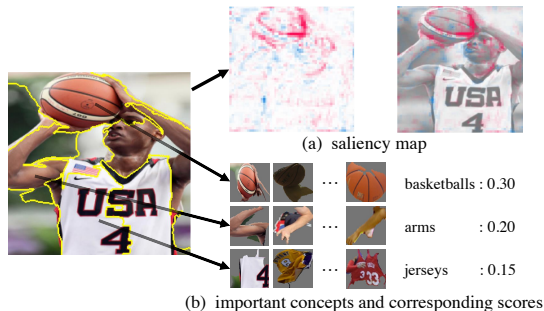
Jiahui Li<sup>1</sup>, Kun Kuang<sup>1\*</sup>, Lin Li<sup>1</sup>, Long Chen<sup>2†</sup>, Songyang Zhang<sup>3</sup>, Jian Shao<sup>1</sup>, Jun Xiao<sup>1</sup>. 2021. Instance-wise or Class-wise? A Tale of Neighbor Shapley for Concept-based Explanation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475337>

## 1 INTRODUCTION

Deep neural networks have demonstrated remarkable performance in many data-driven and prediction-oriented applications [19, 21, 47], and sometimes even perform better than humans. However, their most significant drawback is the lack of interpretability, which makes them less attractive in many real-world applications. When relating to the moral problem or the environmental factors that are uncertain such as crime judgment [10], financial analysis [5], and medical diagnosis [37], it is essential to mine the evidence for the model's prediction (interpret model knowledge) to convince humans. Thus, investigating how to interpret model knowledge is of paramount importance for both academic research and real applications.

The mainstream approaches to interpret model knowledge are *feature-based* methods [2, 4, 9, 24, 26, 28, 29, 31–33, 36, 39, 41, 46], which provide **instance-wise** explanation. They allocate the importance scores for each individual feature in each instance (e.g., each pixel in an image). Based on these importance scores, a saliency map that reflects the accordance for a model's decision intuitively for each instance is provided to improve humans' trusts. For example, in Figure 1(a), those red areas indicate the pixels which contribute most to the model to classify the image as basketball. However, those feature-based interpretations are not consistent with human understanding [23], hence cannot help more on human decision and inference. Humans understand an image always based on high-level concepts, such as segments of basketballs, arms and jerseys as shown in Figure 1(b), rather than low-level pixels.

To bridge the gap between human understanding and model interpreting, some *concept-based* methods [15, 18, 20, 23, 42, 43, 45] which provide **class-wise** explanation have been proposed recently. A concept can be a color, texture, or a group of similar segments that is easy for humans to understand. These methods focus on mining a set of meaningful and representative concepts for an explained class and assigning importance scores according to their contributions to this class. Then, prototypes of the most important concepts found by the model are enumerated to convince humans. However, all of the existing concept-based methods ignore differentiating the importance of concepts on each instance in a class. For example, steering wheel is a key concept for a model to recognize a car,



**Figure 1: Interpreting the knowledge of a classification model to recognize an image of basketball by (a) feature-based explanation and (b) concept-based explanation.**

these methods assigned the same importance score for steering wheel on all images, even for those without steering wheel visually. Hence, they lack the capacity for the instance-wise interpretation of model knowledge. Meanwhile, these methods regard each concept as an independent component, ignoring the interactions among concepts. For example, guitar’s string is an important concept, but without the participation of the guitar’s body, the model can not distinguish whether the object in the image is a guitar or a violin. Thus, it is not appropriate to calculate the contributions for each individual without considering its collaborators.

Hence, we are still facing the following challenges in interpreting model knowledge based on the concept to increase human understanding and trust in models: (i) **Class-wise and instance-wise explanations.** Class-wise explanations interpret the decision boundary of the model, while instance-wise explanations show the unique importance of concepts for each instance. Both of them are important and necessary to increase human understanding and trust in models. (ii) **Interactions among concepts.** Concepts are far from independent, they could be physically interacted (each segment cooperates with its adjacent areas) or semantically interacted (each segment cooperates with its semantically similar areas). (iii) **Evaluation of concept.** As stated in ACE [15], a good concept should satisfy the properties of meaningfulness, coherency, and importance, but how to quantify those properties of discovered concepts is still a problem.

To address these challenges in interpreting model knowledge based on the concept for increasing human understanding and trust in models, we first propose a novel CONE-based NEighbor Shapley (dubbed as CONE-SHAP) method to approximate the importance of each concept with considering its interactions with the physical and semantic neighbors. Meanwhile, we interpret the model knowledge from both instance-wise and class-wise, *i.e.*, we calculate the neighbor Shapley Value of each possible concept in each instance and the top-ranked concepts give an explanation on the instance level (instance-wise) and average the importance of concepts over all instances in a class and selecting top-ranked concepts for explaining on the class level (class-wise). Finally, to comprehensively quantify the discovered concepts and their importance score for interpreting model knowledge, we propose three criteria: coherency, complexity, and faithfulness. We validate our CONE-SHAP with extensive experiments. The results demonstrate

that our algorithm outperforms both feature-based and concepts-based methods in interpreting model knowledge.

The main contribution we made in this paper can be summarized as follows:

- We investigate the problem of how to interpret model knowledge with concept-based explanations from both instance-wise and class-wise.
- We propose a novel CONE-SHAP method to approximate the Shapley Value of each segment with considering its physical and semantic neighbors
- We propose three criteria (coherency, complexity, and faithfulness) to comprehensively quantify the quality of the discovered concepts and their importance scores.
- Extensive experiments show our CONE-SHAP outperforms the existing feature-based and concept-based methods on both class-wise and instance-wise explanations on models.

## 2 RELATED WORK

### 2.1 Feature-based Explanation

Feature-based explanation methods focus on assigning importance scores for the features in an instance (*e.g.*, pixels for an image, words for a text). These methods can be further categorized into several branches: (i) Perturbation-based methods [2, 13, 44, 48], they quantify each feature by measuring the variation of outputs when masking or disturbing that feature while keeping the remaining fixed; (ii) Backpropagation-based methods [3, 29, 39], they compute the importance scores of all features through a few times gradient-related operations; (iii) Model-based methods [32, 33], they employ an explainable model to interpret the original model locally or train an extra deep network, which can output the feature importance directly. These methods can visualize the important features in each instance for explanation (*e.g.*, highlight the pixels in an image). But these feature-based explanations are always not consistent with human understanding [23]. Moreover, these methods above assume all the features are independent and underestimate the interactions among features.

### 2.2 Concept-based Explanation

Humans always understand an object based on high-level concepts rather than fine-grained features. Concepts are defined as a group of similar prototypes which can be understood easily by humans. Given a human-defined concept, Been *et al.* [23] proposes TCAV to quantify the significance of the concepts in a class different from other categories based on trained linear classifiers. ACE [15] employs super-pixel segmentation and cluster methods for mining concepts automatically, and then adopted TCAV [23] for the concept-based explanation. ConceptSHAP [43] defines the notion of completeness score to measure the semantic expression ability of a concept and utilized Shapley Value [34] to find a complete set of concepts. All of these concept-based methods explain deep models on the class-level, but underestimate the local structure of each instance.

## 2.3 Shapley Value for Explaining Models

Shapley Value [25, 34, 35] originated from cooperative game theory and is the best way to distribute benefits fairly by considering the contributions of various agents. Thus, some of the recent studies borrow the idea from Shapley Value to interpret deep neural network [2, 7, 16]. Nevertheless, its computational complexity increases exponentially with the number of participating members. Since there may exist a large number of features in an explained instance, it is exorbitant for a computer to calculate the Shapley Value. To overcome this challenge, the approximation of Shapley Value [6, 12, 22, 27, 30], is used as a substitution which might slightly break some properties of Shapley Value. Ghorbani *et al.* [17] utilize a sample-based method to estimate the Shapley Value for each neuron in the model. Ancona *et al.* [2] also adopt a sample-based method to approximate the Shapley Value and allocate the importance for all features via one-time forward propagation. Chen *et al.* [7] propose L-Shapley and C-Shapley to approximate Shapley value in a graph structure. Though these approximations have made excellent performance in many tasks, most of them focus on features and play roles on the instance level.

Our method overcomes the above shortcomings and explains the model knowledge from both class-level and instance-level. The details will be demonstrated in *Section 5*.

## 3 PRELIMINARIES OF SHAPLEY VALUE

Shapley Value [7, 18, 34] is designed in cooperative game theory to distribute gains fairly by considering the contribution of several players working in a big coalition. Assume a game consists of  $N$  players and they cooperate with each other to achieve a common goal. Let  $u(\cdot)$  represents the utility function to measure the contributions made by an arbitrary set of players. For a particular player  $i$ , let  $S$  be an arbitrary set that contains player  $i$ , and  $S \setminus \{i\}$  represents the set with the absence of  $i$ , then the marginal contribution of  $i$  in  $S$  is defined as:

$$m(i, S) = u(S) - u(S \setminus \{i\}). \quad (1)$$

The Shapley Value of player  $i$  is defined as:

$$\phi_v(i) = \frac{1}{N} \sum_{k=1}^N \frac{1}{\binom{N-1}{k-1}} \sum_{S \in S_k(i)} m(i, S), \quad (2)$$

where  $S_k(i)$  denotes a set with size  $k$  that contain the player  $i$ . Shapley Value is the unique value to satisfy the following properties:

- **Efficiency:** The value of the whole union  $v(\{1, \dots, N\}) - v(\phi)$  is equal to the sum of the Shapley Values of all of the players  $\sum_{i=1}^N \phi_v(i)$ .
- **Symmetry:** If  $\Delta v(i, S) = \Delta v(j, S)$  for all subsets  $S$  then  $\phi_v(i) = \phi_v(j)$ .
- **Dummy:** If  $\Delta v(i, S) = 0$  for all subsets  $S$  then  $\phi_v(i) = 0$ .
- **Additivity:** Let  $u$  and  $w$  represent the associated utility functions, then  $\phi_{v+w}(i) = \phi_v(i) + \phi_w(i)$  for every  $i \in N$ .
- **Coherency:** Given another value function  $\Delta v'(i)$  to measure the marginal contribution of  $i$ , if  $\Delta v(i, S) \geq \Delta v'(i, S)$  for all subsets  $S$ , then  $\phi_v(i) \geq \phi_v'(i)$ .

## 4 EVALUATING CONCEPTS

### 4.1 Problem Formulation

Given a trained neural network  $f$ , and a set of input examples  $X^k = \{x_1^k, x_2^k, \dots, x_n^k\}$  to be explained for a target class  $k$ , where  $x_i^k$  denote the  $i^{th}$  example of class  $k$ , concept-based methods aim at finding out a set of meaningful concepts  $C^k = \{C_1^k, C_2^k, \dots, C_m^k\}$  as well as importance scores  $SC^k = \{SC_1^k, SC_2^k, \dots, SC_m^k\}$  for each concept according to its contribution to the model for class  $k$ . For convenience, we omit the superscript of the above symbols and rewrite  $X^k$  as  $X = \{x_1, x_2, \dots, x_n\}$ ,  $C^k$  as  $C = \{C_1, C_2, \dots, C_m\}$  and  $SC^k$  as  $SC = \{SC_1, SC_2, \dots, SC_m\}$  without causing ambiguity.

### 4.2 Criteria for Evaluating Concept Score

**Previous Methods and Shortcomings.** ACE [15] utilizes *smallest sufficient concepts* (SSC) and *smallest destroying concepts* (SDC) to quantify the quality of the concepts. SSC means to find a set of concepts which are enough for the model to make a prediction, and it is used to measure the representation ability of the extracted concepts. SDC means to look for a set of concepts that will cause a poor prediction when these concepts are removed, and it reflects the necessity of the concepts for a model's decision. ConceptSHAP [43] proposes *completeness* to measure the expression ability of concepts. These existing metrics for evaluating concepts only consider part of the concepts' properties for a model's decision and lack the capacity to measure the concepts from different aspects. Thus, in this paper, we quantify three criteria (coherency, complexity, and faithfulness) to comprehensively evaluate concepts.

**High Coherency.** A concept with a higher score should have a stronger ability to express the semantic of its original inputs, we define this property as *high coherency*. Let  $f$  be a pre-trained model, which maps an input  $x$  to an output  $f(x)$ . And let  $h(\cdot)$  be the function that maps input  $x$  to a representation layer. We define the similarity between the  $i^{th}$  concept and its original input as:

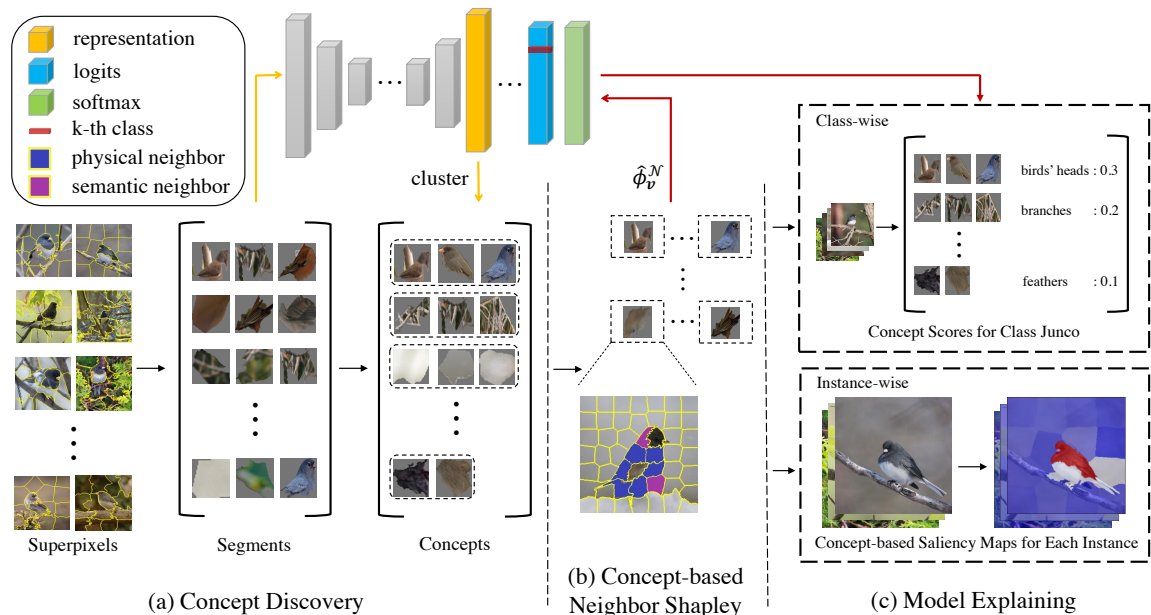
$$\eta(C_i; X) = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} \text{Sim}(h(x|c_{i,j}), h(c_{i,j})), \quad (3)$$

where  $|C_i|$  denotes the numbers of segments in the inputs which belongs to concept  $i$ ,  $c_{i,j}$  denotes the  $j^{th}$  segment of the  $i^{th}$  concept,  $x|c_{i,j}$  denotes the original input instance which contains  $c_{i,j}$  and  $\text{Sim}(\cdot, \cdot)$  denotes the function which measure the similarity between two tensor such as cosine similarity. Assuming that  $\{1 : k\}$  represent the first  $k$  value of a variable, thus  $C\{1 : k\}$  was the top- $k$  concepts with the highest contribution, and  $SC\{1 : k\}$  are their corresponding importance scores, the top- $k$  coherency is defined as:

$$\zeta_k(SC, \eta; C, X) = \text{corr}(SC\{1 : k\}, \eta(C; x)\{1 : k\}), \quad (4)$$

where  $\text{corr}(\cdot, \cdot)$  represents correlation coefficient between two variables.

**Low Complexity.** We want the distribution of the scores of different concepts should be distinguished from each other as much as possible. The concepts' scores are considered complex if they are the same or very close to each other. First, we normalize the scores of top- $k$  concept, and the normalized concept score of the



**Figure 2: The framework of our CONE-SHAP. (a) Segmenting all the samples via super-pixel operation and then cluster the segments into concepts according to their representations in the explained model. (b) Calculating the importance scores for the concepts in each instance and allocate the concepts' scores for an explained class by averaging the importance of concepts over all instances in this class. (c) Providing instance-wise explanation via concept-based saliency maps and provide class-wise explanation via concepts with importance scores.**

$i^{th}$  concept is defined as:

$$\tilde{S}_{C_i} = \frac{SC_i}{\sum_{j=1}^k SC_j}. \quad (5)$$

Note that  $\tilde{S}_{C_i}$  can also be treated as valid probability distribution, then we define the top- $k$  complexity of the concepts as the entropy of  $\hat{S}_i$ :

$$\xi_k(\tilde{S}_{C_i}) = -\sum_{i=1}^k \tilde{S}_{C_i} \ln \tilde{S}_{C_i}. \quad (6)$$

**High Faithfulness.** The change of the model's outputs when the concepts are removed or set to baseline should be correlated with the concepts' scores, we define this property as *high faithfulness*. Let  $X \setminus \{C_i\} = \{x_1 \setminus \{C_i\}, x_2 \setminus \{C_i\}, \dots, x_n \setminus \{C_i\}\}$  denotes the input samples with the absence of all the segments belongs to the  $i^{th}$  concept, where  $x_j \setminus \{C_i\}$  denotes the  $j^{th}$  processed input instance by setting all the segments in the  $i^{th}$  concept as zero or baseline. Let:

$$\varphi(C_i; X) = \frac{1}{n} \sum_{j=1}^n (f(x_j) - f(x_j \setminus C_i)), \quad (7)$$

where  $\varphi(C_i; X)$  represents the degree of the degradation of the model's performance. Since the number of the segments of the concepts various, we normalize  $\varphi(C_i; X)$  as:

$$\tilde{\varphi}(C_i; X) = \frac{1}{|C_i|} \varphi(C_i; X). \quad (8)$$

Then the faithfulness of the top- $k$  concepts' scores is defined as:

$$\theta_k(SC, \varphi; C, X) = \text{corr}(SC\{1:k\}, \tilde{\varphi}(C; X)\{1:k\}). \quad (9)$$

## 5 EXPLAINING MODEL VIA CONE-SHAP

To bridge the gap between model decision and human understanding, we propose a post-hoc approach, namely CONE-SHAP, to interpret the decision procedures of any deep neural network from both instance-wise and class-wise levels with human-friendly concepts. The framework of our CONE-SHAP is exhibited in Figure 2. In order to interpret model knowledge for a target class, our method first extracts concepts automatically and then computes an importance score for each concept according to our proposed CONE-SHAP. Finally, the concept-based saliency maps of each instance are given for instance-wise explanation, and the concepts' contributions for the whole class are given for class-wise explanation.

### 5.1 Concept Discovery

Concepts are defined as prototypes that are understandable for humans [42, 43]. In computer vision tasks, it can be a color, texture, or a group of similar segments, and in natural languages, it can be a group of sub-sentences with the same meaning. Since there are no user-defined concepts in real-world applications, a method to discover concepts automatically is needed. Following ACE [15], we assume a set of semantically similar images' segments as a visual concept. To collect such kind of concept, firstly, a super-pixel method is performed on each sample of the inputs  $X$ , thus we get a dozen segments from a particular class. Then, these segments are clustered into  $m$  different concepts according to their representations computed by model  $f$ . We denote the concepts as  $C = \{C_1, C_2, \dots, C_m\}$ , where  $C_i = \{c_{i,1}, c_{i,2}, \dots, c_{i,|C_i|}\}$ ,  $c_{i,j}$  denotes

the  $j^{th}$  segment in the  $i^{th}$  concept,  $|C_i|$  is the number of the segments belong to the  $i^{th}$  concept, and  $m$  denotes the number of the discovered concepts.

Notice that such an approach relies on a good super-pixel method because the objects of different sizes occupy different proportions in the image which may keep some meaningful concepts from being discovered. For example, the ‘jersey’ in Fig 1 is divided into two parts, though they are a whole. To avoid the meaningful concepts are missed, the multi-grained super-pixel method with three different sizes (small, medium, large) is adopted thus we get the segments with multi-resolution. The details and the hyperparameters will be discussed in Section 6.

## 5.2 Concept-based Neighbor Shapley

To measure the contribution of a segment in an instance for an explained model, we apply a counterfactual method which considers how the prediction of the model will change if this segment is absent. For classification tasks, let  $g$  be the last layer before the *softmax* operation and  $g_k$  represents the logit values of class  $k$ . The value of a segment  $s$  of class  $k$  for the model is calculated as:

$$v_k(s) = g_k(x) - g_k(x \setminus \{s\}). \quad (10)$$

For convenience, we denote  $v_k(s)$  as  $v(s)$ .

**Shapley Value.** We consider all the  $N$  segments in an image as a union, and each of them is a player. Given a particular player  $i$ , let  $S$  be a subset that contains player  $i$  and  $S \setminus \{i\}$  denotes the subset without the participation of  $i$ , then the contribution of  $i$  to the subset  $S$  is computed as:

$$\Delta v(i, S) = v(S) - v(S \setminus \{i\}). \quad (11)$$

When we treat  $v(\cdot)$  as the utility function, then  $\Delta v(\cdot)$  becomes the marginal contribution of the Shapley value. Thus, the Shapley Value of player  $i$  is the weighted average of marginal contribution in all of the subsets:

$$\phi_v(i) = \frac{1}{N} \sum_{j=1}^N \frac{1}{\binom{N-1}{j-1}} \sum_{S \in S_j(i)} \Delta v(i, S), \quad (12)$$

where  $S_j(i)$  denotes the set with size  $j$  that contains the  $i^{th}$  segment.

However, the computational complexity for Shapley Value increases exponentially as the number of players rises [14, 25, 38]. Since each image contains more than a hundred segments, it is costly for a computer to calculate the truly Shapley Value. Thus, recent studies replaced the truly Shapley Value with its approximation [2, 7, 8, 17] in different situations.

**Approximation of Shapley Value.** Inspired by [14], the players can be treated as the nodes of a fully connected graph, where any two players are connected since they might have correlations during a game. In the application of image classification, the segments of an image can also be treated as nodes, but each node only connects with its neighbors. Here, we define the neighbors  $\mathcal{N}(i)$  of the  $i^{th}$  segment are those segments which are adjacent to it (physical neighbors) or belong to the same concept as it (semantic neighbors). Based on the assumption that participants which are not the neighbors of  $i$  hardly affect its contribution for a model’s inference procedure, the

Shapley Value of  $i$  in Equation 12 can be approximated as:

$$\phi_v^{\mathcal{N}}(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{j=1}^{|\mathcal{N}(i)|} \frac{1}{\binom{|\mathcal{N}(i)|-1}{j-1}} \sum_{\substack{S \in S \\ S \subseteq \mathcal{N}(i)}} \Delta v(i, S). \quad (13)$$

Considering that a segment may contain a large amount of neighbors in an instance, we adopt sample-based method to estimate  $\phi_v^{\mathcal{N}}(i)$  in order to further reduce the computational costs. Concretely, we first sample  $k$  nodes from  $\mathcal{N}(i)$  and denote it as  $\mathcal{N}_k(i)$ , and then compute the Shapley Value in the  $\mathcal{N}_k(i)$ . This procedure will repeat  $M$  times, and we take the average of these results as the COncept-based NEighbor Shapley Value (CONE-SHAP) of  $i$ :

$$\hat{\phi}_v^{\mathcal{N}}(i) = \frac{1}{M|\mathcal{N}_k(i)|} \sum_{t=1}^M \sum_{j=1}^{|\mathcal{N}_k(i)|} \frac{1}{\binom{|\mathcal{N}_k(i)|-1}{j-1}} \sum_{\substack{S \in S \\ S \subseteq \mathcal{N}_k(i)}} \Delta v(i, S). \quad (14)$$

Next, we will introduce how to employ the approximation of Shapley Value from Equation 14 to interpret model knowledge from both instance-wise and class-wise. And more discussion between CONE-SHAP and other approximations of Shapley Value can be found in Appendix.

## 5.3 Model Explaining

**Instance-wise Explanation.** In order to help users understand the basis for a model’s reasoning procedure intuitively, we provide concept-based saliency maps to interpret model knowledge on the instance-level. The contribution of each segment of each instance is assigned according to its CONE-SHAP Value  $\hat{\phi}_v^{\mathcal{N}}(i)$ . Compared to perturbation-based methods which explain a model in fine-grained features, our CONE-SHAP focuses on the concept-based explanation, which is more human-friendly.

**Class-wise Explanation.** To interpret model knowledge on the class level, our method distributes the concept scores to indicate which concept contributes more to the model’s prediction on the explained class. A concept is considered important if all of its belongings own a high Shapley Value. Since we have gotten a group of possible concepts in the concept discovery procedure, for a concept  $C_i$ , we compute its score by averaging all of the approximate Shapley Values of its segments:

$$SC_i = \frac{1}{|C_i|} \sum_{c_{i,j} \in C_i} \hat{\phi}_v^{\mathcal{N}}(i). \quad (15)$$

## 5.4 Analysis on CONE-SHAP

**Analysis of Computational Complexity.** The Shapley Value of a segment is computed according to Equation 12. When we interpret an target instance, let  $n$  be the average number of segments per instance, then the computational complexity of Equation 12 is  $O(2^n)$ . When we approximate Equation 12 via Equation 13 which assume that the participated players only cooperate with its neighbors, the computational complexity downgrade to  $O(2^{|\mathcal{N}|})$ .  $\mathcal{N}$  is the average number of the neighbors for a segment which is much more smaller than  $n$ . When  $\mathcal{N}$  is large, our proposed CONE-SHAP estimates Equation 13 via Equation 14, then the computational complexity becomes  $O(M2^k)$ . Since both  $k$  and  $M$  are small constants, the final computational complexity can be written as  $O(d)$ , where  $d$  is a positive integer.

**Analysis of the Approximation.**

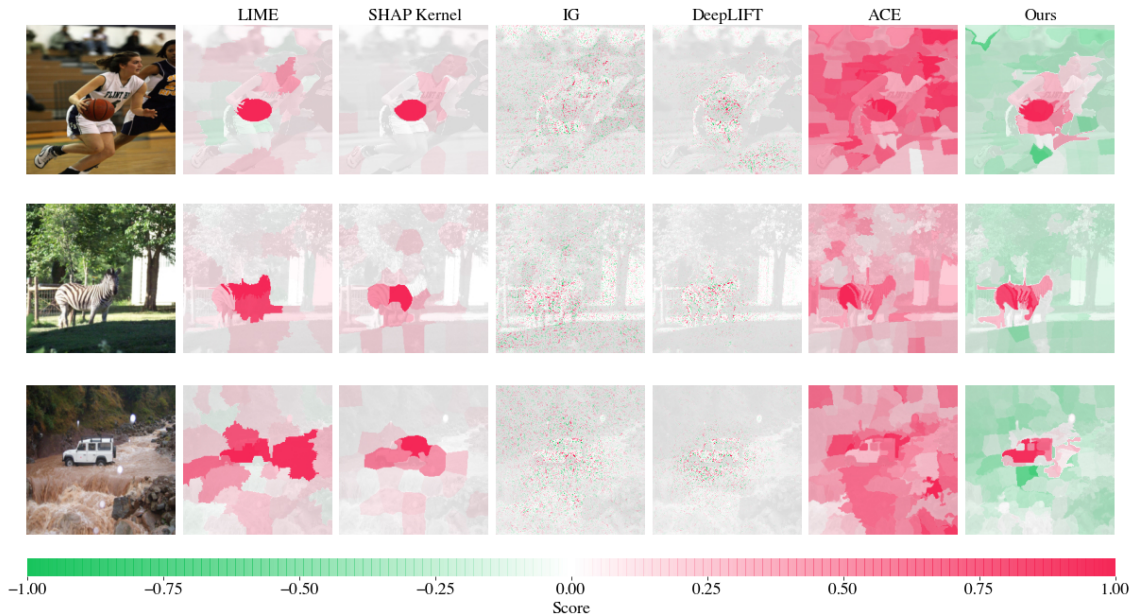


Figure 3: The comparison of the saliency maps of our method and existing popular methods (feature-based and concept-based). Three images in ImageNet are selected as examples and pre-trained densenet-121 is used as the explained model.

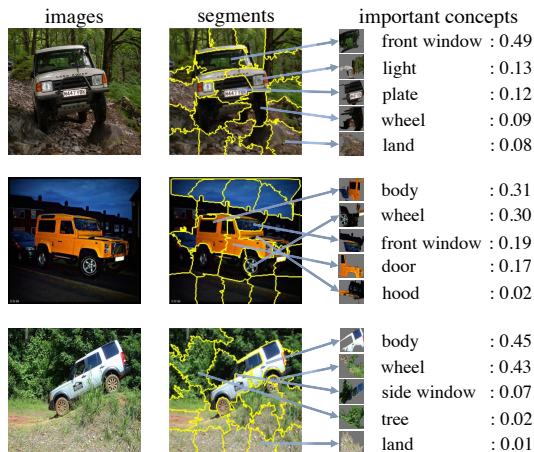


Figure 4: Show cases of class jeep in ImageNet and their 5 most important explained by our method. The same concept might have different importance across instances (images).

**Theorem 1.** When the mutual information between a player  $i$  and any of its neighbors is bounded by  $\epsilon$ , the expected error between Equation 13 and Equation 12 is bounded by  $2\epsilon$ :

$$\mathbb{E}|\phi_v^N(i) - \phi_v(i)| \leq 2\epsilon. \quad (16)$$

**Theorem 2.** Equation 14 converges to Equation 13 as  $M \rightarrow \infty$ .

We leave the proofs in Appendix.

## 6 EXPERIMENTS

In this section, we first discuss the experimental settings and the hyperparameters we adopt. Then, we demonstrate the interpretation ability of our method from both instance-wise and class-wise.

Meanwhile, we evaluate the efficiency of the top-ranking concepts. Finally, the rationality of the allocated concepts' scores is measured according to the criteria that we proposed.

### 6.1 Experimental Settings

Our method can be applied to any task without further training. In order to demonstrate the superiority of our method intuitively, we focus on image classification in this paper.

**Models and Dataset.** We interpret two most commonly used official models Densenet-121 [21] and Inception-V3 [40] which pre-trained on the ImageNet dataset [11] in the experiments. The former is used for instance-wise explanation and the latter is used for class-wise explanation.

**Settings for Concept Discovery.** For each explained instance, we first picked out all of the samples that belong to the same class, and then mine out the candidate concepts for this class by segmenting these instances and using K-means method to cluster the fragments. Since the meaningful concepts for an object are usually no more than 20, we set the cluster center to 20 for each class. As we mentioned before, the discovery of the concepts relies on a good segmentation method, in order to avoid neglecting meaningful segments, we perform multi-resolution super-pixel for each image. Concretely, SLIC [1] is used to segment input samples for speed, and each instance is divided into three granularities with the resolutions of 15, 50, 80 segments separately. Thus, the concepts with different sizes are captured.

**Settings for Feature Extraction.** As for the features we used to cluster the image fragments, we resize the segments to the size of the original images via bicubic interpolation, and feed them to the explained model to get the representations of the middle layer of the neural network. We adopt *ave-pool* layer in Inception-V3 [40]

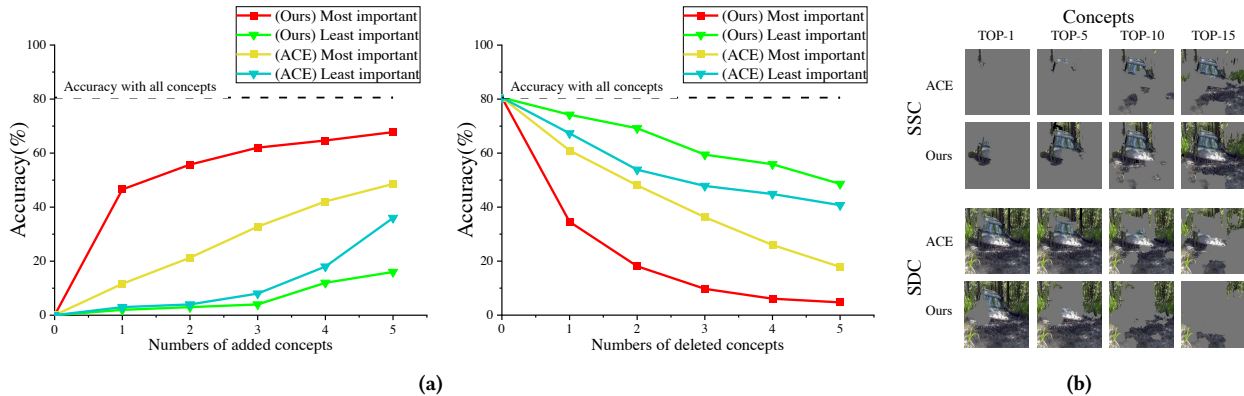


Figure 5: The comparison of performance of the top-ranked concepts computed by our CONE-SHAP and ACE. (a) The changes of the model’s performance when the most important 5 concepts are added/removed gradually. (b) An example of removing/adding the top-ranked concepts of an image in class jeep.

and *global ave-pool* layer in Densenet-121 [21] as the representation layer respectively.

**Settings for Approximated Shapley Value.** For the importance of each segment, we compute the approximated Shapley Value according to Equation 14. To make the balance between the computational complexity and the interpretation performance, we set the hyperparameter  $k$  as 5 and  $M$  as 1 to sample while calculating the approximated Shapley Value. The choice of these hyperparameter will be discussed in Subsection 6.3.

## 6.2 Instance-wise Explanation

For instance-wise explanation, we compare our method to LIME [32], SHAP kernel [29], Integrated Gradients (IG) [39], DeepLIFT [36], and ACE [15] via saliency maps. Specially, ACE is a concept-based method, and can’t provide saliency maps directly, thus we got the saliency maps by assigning the TCAV [23] scores to its segments. We demonstrate ACE here in order to show that the other concept-based methods lack the capability to explain the model from instance-wise.

**Explanation with Saliency Maps.** We provide coarse-grained saliency maps with the same size of the inputs to indicate which part of an instance is more important. Since we have got the approximated Shapley Value of each segment of each image with three different granularities, we treat these values as the scores in saliency maps, and then we get three saliency maps focus on different scales. For each instance, we average the multi-grained saliency maps to get the final saliency map. Our method provides concept-based saliency maps which are coarse-grained but more human-friendly.

Figure 3 shows the comparison of the saliency maps between our CONE-SHAP and baselines, where the importance scores of each unit are scaled between -1 and 1 by dividing the absolute value of the maximum number, and the units with scores below zero were set to green and beyond zero were set to red. There exist more green areas in our method because we consider the correlation between each area adequately, and the backgrounds and the meaningless segments hardly cooperate with the others. Compared with methods that distribute importance scores for each pixel, our concept-based saliency maps are easier for humans to

understand. ACE [15] focus on interpreting concepts from class-wise thus fails to provide precise explanations for each instance.

**Different Importance of Concepts on Different Instances.** Intuitively, even the same concept might have different importance for different instances/images. Based on the value of CONE-SHAP on each segment, our method can estimate the importance of each concept on each instance as follows: Firstly, we find out all of the concepts in an instance. Then, for each concept, we compute the CONE-SHAP value of each segment that belongs to the concept with Equation 14. Finally, we can estimate the importance of the concept by sum up the CONE-SHAP values of its segments.

Figure 4 exemplifies the interpretation of our CONE-SHAP for a model to recognize *jeep*. In the class-level, our method shows the concept of bodies, windows, plates, and wheels gears are very important for the class of jeep, while in a specific instance/image, different concepts have different importance. For example, in the first picture of Figure 4, the most important concept is front window with the score 0.49, but we can not see the front window in the third picture, hence the most important concept in the third image changes to jeep’s body. And wheel is important in all of the three pictures but it plays different roles in different instances.

## 6.3 Class-wise Explanation

The compared baselines<sup>1</sup> for class-wise explanation include ACE [15], SHAP(MC), which approximates the Shapley Value with Monte Carlo sampling, and Occlusion which compute the importance for each segment according to Equation 11.

**Validating the Performance of Concepts.** To measure the top- $k$  important concepts on the explained model, we employ the same metrics (*i.e.*, SSC and SDC) as ACE [15], where SSC/SDC represents the accuracy of model prediction when we add/remove the most important concepts on the image. We use official Inception-V3 [40] without any further training as the explained model. 20 classes in ImageNet are selected randomly to explain, and we calculate SSC and SDC for each class separately and take an average.

Table 1 reports the changes of the prediction accuracy by gradually removing/adding the top- $k$  important concepts, where the

<sup>1</sup>We did not compare with TCAV [23] and ConceptSHAP [43], since they rely on the dataset with human-labeled concepts.

**Table 1: (a) The changes of the prediction accuracy when gradually removing the top- $k$  important concepts (the lower the better). (b) The changes of the prediction accuracy when gradually adding the top- $k$  important concepts (the higher the better).**

(a)						(b)					
Methods	SDC (removing top- $k$ important concepts)					Methods	SSC (adding top- $k$ important concepts)				
	top-1	top-2	top-3	top-4	top-5		top-1	top-2	top-3	top-4	top-5
ACE	60.97	48.07	36.27	25.98	17.88	ACE	11.63	21.28	32.78	42.09	48.57
Oclusion	34.58	19.93	14.10	11.61	8.38	Oclusion	44.40	51.35	55.96	59.67	61.36
SHAP (MC)	34.48	18.58	13.15	10.42	5.91	SHAP (MC)	44.56	52.36	55.43	60.36	63.36
<b>Ours</b>	<b>34.15</b>	18.11	<b>9.75</b>	<b>6.08</b>	<b>4.73</b>	<b>Ours</b>	<b>46.56</b>	<b>55.76</b>	<b>62.01</b>	64.68	<b>67.78</b>
Ours ( $w/o$ PN)	36.56	<b>17.72</b>	12.42	9.44	4.92	Ours ( $w/o$ PN)	44.02	54.86	59.62	63.36	66.00
Ours ( $w/o$ SN)	34.75	19.45	12.29	6.08	4.86	Ours ( $w/o$ SN)	45.56	54.59	60.50	<b>64.90</b>	66.41

**Table 2: (a) The mean changes of the quality of the concepts when the number of sample times  $M$  increases when we classify 5 classes selected randomly in the test set of ImageNet. (b) The mean changes of the quality of the concepts when the number of sampled neighbors  $k$  increases when we classify 5 classes selected randomly in the test set of ImageNet.**

(a)				(b)					
Metrics	M=1	M=2	M=3	Metrics	k=1	k=2	k=3	k=4	k=5
	SSC-most	43.20	44.00		42.00	SSC-most	31.60	34.00	38.80
SSC-least	6.40	6.80	5.20	SSC-least	11.20	8.80	8.00	8.00	5.60
SDC-most	33.20	34.00	33.20	SDC-most	41.20	40.40	36.80	36.40	32.40
SDC-least	82.80	85.20	86.00	SDC-least	83.20	83.80	82.40	84.00	84.80

method  $w/o$  PN and  $w/o$  SN refer to the ablations from our method by removing the Physical Neighbor (PN) and Semantic Neighbor (SN) in Equation 13 & 14, respectively. From the results, we can conclude that (i) By removing/adding the top- $k$  important concepts, our method makes the model achieve the lowest/highest accuracy. This is because our method can estimate the importance of each concepts more precisely than baselines. (ii) Both of the ablations ( $w/o$  PN and  $w/o$  SN) would lead a worse performance on our method, which indicates that both of the physical and semantic neighbor considered in our CONE-SHAP method is necessary. We also plot Figure 5 to clearly demonstrate the advantages of our CONE-SHAP compared with ACE. From the Figure 5(a), we have following observations: (i) By adding/deleting the most important concepts from the image, our method changes (improve/reduce) the model accuracy more remarkably than baseline ACE. (ii) By adding/deleting the least important concepts, our method have less influence on the model accuracy than baseline. Moreover, we show an example of adding/removing the most important concepts of an image in Figure 5(b).

**Analysis of the Hyperparameters for Approximating Shapley Value.** We approximated the truly Shapley Value via sampling from the neighbors as depicted in Equation 14. A large  $M$  and  $k$  will bring pressure to computation costs, and a small  $M$  and  $k$  will lead to inaccurate estimates. We performed extensive experiments to find moderate  $M$  and  $k$ . For choosing  $M$ , we select five classes randomly to interpret. We fix  $k$  to 5 and increase  $M$  gradually to compute *SSC-most*, *SSC-least*, *SDC-most*, and *SDC-least* of the selected classes. The results shown in Table 2(a) demonstrate that in our settings, set  $M$  to 1 is enough to approximate the Shapley Value. Then for choosing  $k$ , we set  $M$  to 1 and increase  $k$  gradually. We find that 5 neighbors is sufficient to reach relatively high performance.

**Table 3: The quality of the allocated scores for the top-5 concepts on the criteria we proposed.**

Methods	Criteria top-5		
	Coherency	Complexity	Faithfulness
ACE	0.4090	1.5943	0.4516
Oclusion	0.7437	1.4051	0.9245
SHAP (MC)	0.7602	1.3930	0.9325
<b>Ours</b>	<b>0.9299</b>	1.2622	<b>0.9542</b>
Ours ( $w/o$ PN)	0.8423	1.3135	0.9427
Ours ( $w/o$ SN)	0.9235	<b>1.2422</b>	0.9462

Table 2(b) shows the mean results of 5 classes selected randomly to exhibit how the performance changes with the growth of  $k$ . Thus, we set  $M$  to 1 and  $k$  to 5 in our experiments.

**Evaluation of Concept Scores.** SSC and SDC merely take into account whether concepts are ranked rationally, and it is also necessary to evaluate the quality of the distributed scores. Hence, we evaluated the scores computed by various methods based on the three criteria we proposed. We adopt Pearson correlation coefficient to compute the coherency in Equation 3 and faithfulness in Equation 9, and the results are depicted in Table 3. As the table shows, our CONE-SHAP obtains the best score on *coherency* and *faithfulness* that reach 0.93 and 0.95 respectively. That means the concept score and expressive ability are highly consistent, and the scores are highly correlated with its contributions to model performance. Besides, the low complexity reflects our scoring mechanism can distinguish concepts from each other very well.



## 7 CONCLUSION

In this paper, we investigate the problem of post-hoc explanation for deep neural networks in a human-friendly way. First, we propose a method named CONE-SHAP which can explain any model from both instance-wise and class-wise without further training. Especially, by considering the interaction neighbors, our CONE-SHAP downgrades the computational complexity of Shapley Value from exponential to polynomial. Since there are no unified metrics to measure the performance of concepts' scores, we proposed three criteria to evaluate the scoring mechanism of concept-based explanation methods. Extensive experiments demonstrate the superior performance of our method. Applying our method to real-world applications, users can easily understand the important shreds of evidence for a model's prediction so as to make a decision confidently.

## ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Project of China (No.2018AAA0101900), the National Natural Science Foundation of China (No. 61625107, U19B2043, 61976185, No. 62006207), Zhejiang Natural Science Foundation (LR19F020002), Key R & D Projects of the Ministry of Science and Technology (No. 2020YFC0832500), Zhejiang Innovation Foundation(2019R52002), the Fundamental Research Funds for the Central Universities and Zhejiang Province Natural Science Foundation (No. LQ21F020020).

## REFERENCES

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI* (2012).
- [2] Marco Ancona, Cengiz Oztireli, and Markus Gross. 2019. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation. In *ICML*.
- [3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* (2015).
- [4] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*. Springer, 63–71.
- [5] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. 2019. Machine learning explainability in finance: an application to default risk analysis. (2019).
- [6] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36, 5 (2009), 1726–1730.
- [7] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. 2018. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In *ICLR*.
- [8] Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. (2020).
- [9] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 598–617.
- [10] Ashley Deeks. 2019. The judicial demand for explainable artificial intelligence. *Columbia Law Review* 119, 7 (2019), 1829–1850.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
- [12] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. 2008. A linear approximation method for the Shapley value. *Artificial Intelligence* 172, 14 (2008), 1673–1699.
- [13] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*.
- [14] Christopher Frye, Ilya Feige, and Colin Rowat. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. (2020).
- [15] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *NeurIPS*.
- [16] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *ICML*.
- [17] Amirata Ghorbani and James Zou. 2020. Neuron Shapley: Discovering the Responsible Neurons. (2020).
- [18] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *arXiv* (2019).
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [20] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. (2017).
- [21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*.
- [22] Michael A Jones and Jennifer M Wilson. 2010. Multilinear extensions and values for multichoice games. *Mathematical Methods of Operations Research* 72, 1 (2010), 145–169.
- [23] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (teav). In *ICML*.
- [24] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Kuang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. 2020. Causal Inference. *Engineering* 6, 3 (2020), 253–263.
- [25] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. (2020).
- [26] Fashen Li, Lian Li, Jianping Yin, Yong Zhang, Qingguo Zhou, and Kun Kuang. 2020. How to Interpret Machine Knowledge. *Engineering* 6, 3 (2020), 218–220.
- [27] Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. 2021. Shapley Counterfactual Credits for Multi-Agent Reinforcement Learning. *KDD* (2021).
- [28] Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, and Quanshi Zhang. 2020. Knowledge consistency between neural networks and beyond. (2020).
- [29] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NeurIPS*.
- [30] Guillermo Owen. 1972. Multilinear extensions of games. *Management Science* 18, 5-part-2 (1972), 64–79.
- [31] Yunhe Pan. 2020. Multiple knowledge representation of artificial intelligence. *Engineering* 6, 3 (2020), 216–217.
- [32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *SIGKDD*.
- [33] Patrick Schwab and Walter Karlen. 2019. CXPlain: Causal explanations for model interpretation under uncertainty. In *NeurIPS*.
- [34] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* (1953).
- [35] Lloyd S Shapley, Alvin E Roth, et al. 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.
- [36] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *ICML*.
- [37] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. 2020. Explainable deep learning models in medical image analysis. *Journal of Imaging* 6, 6 (2020), 52.
- [38] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. *ICML* (2020).
- [39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2016. Gradients of counterfactuals. *arXiv* (2016).
- [40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.
- [41] Pei Wang and Nuno Vasconcelos. 2020. SCOUT: Self-aware Discriminant Counterfactual Explanations. In *CVPR*.
- [42] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. 2020. Towards Global Explanations of Convolutional Neural Networks With Concept Attribution. In *CVPR*.
- [43] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. 2020. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. *NeurIPS* (2020).
- [44] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*. Springer.
- [45] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *ECCV*.
- [46] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. 2020. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering* 6, 3 (2020), 310–345.
- [47] Yueting Zhuang, Ming Cai, Xuelong Li, Xiangang Luo, Qiang Yang, and Fei Wu. 2020. The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI. *Engineering* 6, 3 (2020), 245.
- [48] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. (2017).