

EXPLAINING ALPHAGO: INTERPRETING CONTEXTUAL EFFECTS IN NEURAL NETWORKS

Anonymous authors

Paper under double-blind review

ABSTRACT

In this paper, we propose to disentangle and interpret contextual effects that are encoded in a pre-trained deep neural network. We use our method to explain the gaming strategy of the alphaGo Zero model. Unlike previous studies that visualized image appearances corresponding to the network output or a neural activation only from a global perspective, our research aims to clarify how a certain input unit (dimension) collaborates with other units (dimensions) to constitute inference patterns of the neural network and thus contribute to the network output. The analysis of local contextual effects *w.r.t.* certain input units is of special values in real applications. Explaining the logic of the alphaGo Zero model is a typical application. In experiments, our method successfully disentangled the rationale of each move during the Go game.

1 INTRODUCTION

Interpreting the decision-making logic hidden inside neural networks is an emerging research direction in recent years. The visualization of neural networks and the extraction of pixel-level input-output correlations are two typical methodologies. However, previous studies usually interpret the knowledge inside a pre-trained neural network from a global perspective. For example, (Ribeiro et al., 2016; Lundberg & Lee, 2017; Kindermans et al., 2018) mined input units (dimensions or pixels) that the network output is sensitive to; (Bau et al., 2017) visualized receptive fields of filters in intermediate layers; (Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015; Simonyan et al., 2013; Dosovitskiy & Brox, 2016; Fong & Vedaldi, 2017; Selvaraju et al., 2017) illustrated image appearances that maximized the score of the network output, a filter’s response, or a certain activation unit in a feature map.

However, instead of visualizing the entire appearance that is responsible for a network output or an activation unit, we are more interested in the following questions.

- How does a local input unit contribute to the network output? Here, we can vectorize the input of the network into a high-dimensional vector, and we treat each dimension as a specific “unit” without ambiguity. As we know, a single input unit is usually not informative enough to make independent contributions to the network output. Thus, we need to clarify which other input units the target input unit collaborates with to constitute inference patterns of the neural network, so as to pass information to high layers.
- Can we quantitatively measure the significance of above contextual collaborations between the target input unit and its neighboring units?

Method: Therefore, given a pre-trained convolutional neural network (CNN), we propose to disentangle contextual effects *w.r.t.* certain input units.

As shown in Fig. 1, we design two methods to interpret contextual collaborations at different scales, which are agnostic to the structure of CNNs. The first method estimates a rough region of contextual collaborations, *i.e.* clarifying whether the target input unit mainly collaborates with a few neighboring units or most units of the input. This method distills knowledge from the pre-trained network into a mixture of local models (see Fig. 2), where each model encodes contextual collaborations within a specific input region to make predictions. We hope that the knowledge-distillation strategy can help people determine quantitative contributions from different regions. Then, given a model for

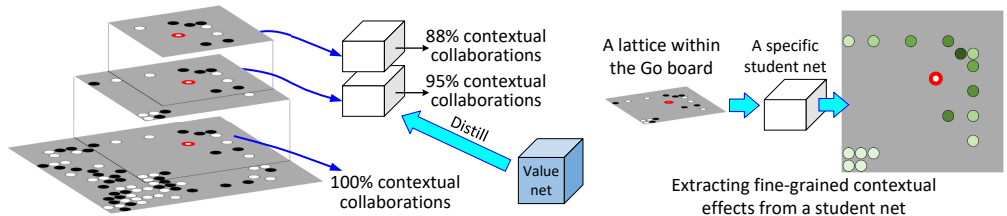


Figure 1: Explaining the alphaGo model. Given the state of the Go board and the next move, we use the alphaGo model to explain the rationale of the move. We first estimate a rough region of contextual collaborations *w.r.t.* the current move by distilling knowledge from the value net to student nets that receive different regions of the Go board as inputs. Then, given a student net, we analyze fine-grained contextual collaborations within its region of the Go board. In this figure, we use a board state from a real Go game between humans for clarity.

local collaborations, the second method further analyzes the significance of detailed collaborations between each pair of input units, when we use the local model to make predictions on an image.

Application, explaining the alphaGo Zero model: The quantitative analysis of contextual collaborations *w.r.t.* a local input unit is of special values in some tasks. For example, explaining the alphaGo model (Silver et al., 2016; Hassabis & Silver, 18 October 2017) is a typical application.

The alphaGo model contains a value network to evaluate the current state of the game—a high output score indicates a high probability of winning. As we know, the contribution of a single move (*i.e.* placing a new stone on the Go board) to the output score during the game depends on contextual shapes on the Go board. Thus, disentangling explicit contextual collaborations that contribute to the output of the value network is important to understand the logic of each new move hidden in the alphaGo model.

More crucially, in this study, we explain the alphaGo Zero model (Hassabis & Silver, 18 October 2017), which extends the scope of interests of this study from *diagnosing feature representations of a neural network* to a more appealing issue *letting self-improving AI teach people new knowledge*. The alphaGo Zero model is pre-trained via self-play without receiving any prior knowledge from human experience as supervision. In this way, all extracted contextual collaborations represent the automatically learned intelligence, rather than human knowledge.

As demonstrated in well-known Go competitions between the alphaGo and human players (alp, Retrieved 17 March 2016; 2017-05-27), the automatically learned model sometimes made decisions that could not be explained by existing gaming principles. The visualization of contextual collaborations may provide new knowledge beyond people’s current understanding of the Go game.

Contributions of this paper can be summarized as follows.

- (i) In this paper, we focus on a new problem, *i.e.* visualizing local contextual effects in the decision-making of a pre-trained neural network *w.r.t.* a certain input unit.
- (ii) We propose two new methods to extract contextual effects via diagnosing feature representations and knowledge distillation.
- (iii) We have combined two proposed methods to explain the alphaGo Zero model, and experimental results have demonstrated the effectiveness of our methods.

2 RELATED WORK

Understanding feature representations inside neural networks is an emerging research direction in recent years. Related studies include 1) the visualization and diagnosis of network features, 2) disentangling or distilling network feature representations into interpretable models, and 3) learning neural networks with disentangled and interpretable features in intermediate layers.

Network visualization: Instead of analyzing network features from a global view (Wolchover, 2017; Schwartz-Ziv & Tishby, 2017; Rauber et al., 2016), (Bau et al., 2017) defined six types of semantics for middle-layer feature maps of a CNN, *i.e.* *objects, parts, scenes, textures, materials,* and *colors*. Usually, each filter encodes a mixture of different semantics, thus difficult to explain.

Visualization of filters in intermediate layers is the most direct method to analyze the knowledge hidden inside a neural network. (Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015; Simonyan et al., 2013; Dosovitskiy & Brox, 2016; Yosinski et al., 2015; Dong et al., 2017; Zhang et al., 2016) showed the appearance that maximized the score of a given unit. (Dosovitskiy & Brox, 2016) used up-convolutional nets to invert CNN feature maps to their corresponding images.

Pattern retrieval: Some studies retrieved certain units from intermediate layers of CNNs that were related to certain semantics, although the relationship between a certain semantics and each neural unit was usually convincing enough. People usually parallel the retrieved units similar to conventional mid-level features (Singh et al., 2012) of images. (Zhou et al., 2015; 2016) selected units from feature maps to describe “scenes”. (Simon & Rodner, 2015) discovered objects from feature maps.

Model diagnosis and distillation: Model-diagnosis methods, such as the LIME (Ribeiro et al., 2016), the SHAP (Lundberg & Lee, 2017), influence functions (Koh & Liang, 2017), gradient-based visualization methods (Fong & Vedaldi, 2017; Selvaraju et al., 2017), and (Kumar et al., 2017) extracted image regions that were responsible for network outputs. (Vaughan et al., 2018; Zhang et al., 2018b) distilled knowledge from a pre-trained neural network into explainable models to interpret the logic of the target network. Such distillation-based network explanation is related to the first method proposed in this paper. However, unlike previous studies distilling knowledge into explicit visual concepts, our using distillation to disentangle local contextual effects has not been explored in previous studies.

Learning interpretable representations: A new trend is to learn networks with meaningful feature representations in intermediate layers (Hu et al., 2016; Stone et al., 2017; Liao et al., 2016) in a weakly-supervised or unsupervised manner. For example, capsule nets (Sabour et al., 2017) and interpretable RCNN (Wu et al., 2017) learned interpretable middle-layer features. InfoGAN (Chen et al., 2016) and β -VAE (Higgins et al., 2017) learned meaningful input codes of generative networks. (Zhang et al., 2018a) developed a loss to push each middle-layer filter towards the representation of a specific object part during the learning process without given part annotations.

All above related studies mainly focused on semantic meanings of a filter, an activation unit, a network output. In contrast, our work first analyzes quantitative contextual effects *w.r.t.* a specific input unit during the inference process. Clarifying explicit mechanisms of how an input unit contributes to the network output has special values in applications.

3 ALGORITHM

In the following two subsections, we will introduce two methods that extract contextual collaborations *w.r.t.* a certain input unit from a CNN at different scales. Then, we will introduce the application that uses the proposed methods to explain the alphaGo Zero model.

3.1 DETERMINING THE REGION OF CONTEXTUAL COLLABORATIONS *w.r.t.* AN INPUT UNIT

Since the input feature usually has a huge number of dimensions (units), it is difficult to accurately discover a few input units that collaborate with a target input unit. Therefore, it is important to first approximate the rough region of contextual collaborations before the unit-level analysis of contextual collaborations, *i.e.* clarifying in which regions contextual collaborations are contained.

Given a pre-trained neural network, an input sample, and a target unit of the sample, we propose a method that uses knowledge distillation to determine the region of contextual collaborations *w.r.t.* the target input unit. Let $I \in \mathbf{I}$ denote the input feature (*e.g.* an image or the state in a Go board). Note that input features of most CNNs can be represented as a tensor $I \in \mathbb{R}^{H \times W \times D}$, where H and W indicate the height of the width of the input, respectively; D is the channel number. We clip different lattices (regions) $\Lambda_1, \Lambda_2, \dots, \Lambda_N \in \mathbf{\Lambda}$ from the input tensor, and input units within the i -th lattice are given as $I_{\Lambda_i} \in \mathbb{R}^{h \times w \times D}$, $h \leq H, w \leq W$. Different lattices overlap with each other.

The core idea is that we use a mixture of models to approximate the function of the given pre-trained neural network (namely the *teacher net*), where each model is a *student net* and uses input information within a specific lattice I_{Λ_i} to make predictions.

$$\hat{y} \approx \alpha_1 \cdot y_1 + \alpha_2 \cdot y_2 + \dots + \alpha_n \cdot y_n \quad (1)$$

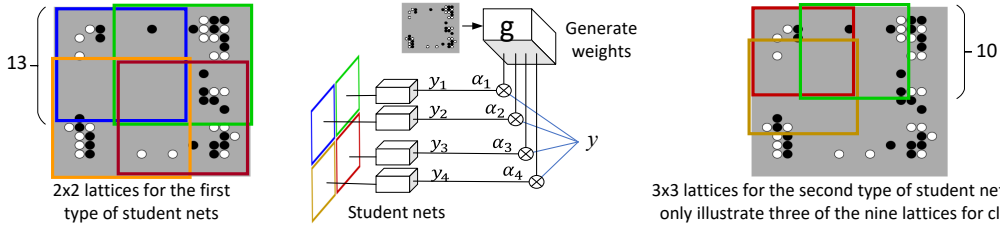


Figure 2: Division of lattices for two types of student nets. We distill knowledge from the value net into a mixture of four/nine student nets to approximate decision-making logic of the value net.

where $\hat{y} = f(I)$ and $y_i = f_i(I_{\Lambda_i})$ denote the output of the pre-trained teacher net f and the output of the i -th student net f_i , respectively. α_i is a scalar weight, which depends on the input I . Because different lattices within the input are not equally informative *w.r.t.* the target task, input units within different lattices make different contributions to final network output.

More crucially, given different inputs, the importance for the same lattice may also change. For example, as shown in (Selvaraju et al., 2017), the head appearance is the dominating feature in the classification of animal categories. Thus, if a lattice corresponds to the head, then this lattice will contribute more than other lattices, thereby having a large weight α_i . Therefore, our method estimates a specific weight α_i for each input I , *i.e.* α_i is formulated as a function of I (which will be introduced later).

Significance of contextual collaborations: Based on the above equation, the significance of contextual collaborations within each lattice Λ_i *w.r.t.* an input unit can be measured as s_i .

$$\Delta\hat{y} \approx \underbrace{\alpha_1 \cdot \Delta y_1}_{\text{Impacts from the first lattice } \Lambda_1} + \alpha_2 \cdot \Delta y_2 + \dots + \alpha_n \cdot \Delta y_n, \quad s_i = |\alpha_i \cdot \Delta y_i| \quad (2)$$

where we revise the value of the target unit in the input and check the change of network outputs, $\Delta\hat{y} = f(I^{\text{new}}) - f(I)$ and $\Delta y_i = f_i(I_{\Lambda_i}^{\text{new}}) - f_i(I_{\Lambda_i})$. If contextual collaborations *w.r.t.* the target unit mainly localize within the i -th lattice Λ_i , then $\alpha_i \cdot \Delta y_i$ can be expected to contribute the most to the change of \hat{y} .

We conduct two knowledge-distillation processes to learn student nets and a model of determining $\{\alpha_i\}$, respectively.

Student nets: The first process distills knowledge from the teacher net to each student net f_i with parameters θ_i based on the distillation loss $\min_{\theta_i} \sum_{I \in \mathbf{I}} \|y_{I,i} - \hat{y}_I\|^2$, where the subscript I indicates the output for the input I . Considering that Λ_i only contains partial information of I , we do not expect $y_{I,i}$ to reconstruct \hat{y}_I without any errors.

Distilling knowledge to weights: Then, the second distillation process estimates a set of weights $\alpha = [\alpha_{I,1}, \alpha_{I,2}, \dots, \alpha_{I,n}]$ for each specific input I . We use the following loss to learn another neural network g with parameters θ_g to infer the weight.

$$\alpha = g(I), \quad \min_{\theta_g} \text{Loss}(\theta_g), \quad \text{Loss}(\theta_g) = \sum_{I \in \mathbf{I}} \left\| \Delta\hat{y}_I - \sum_{i=1}^N \alpha_{I,i} \cdot \Delta y_{I,i} \right\|^2 \quad (3)$$

3.2 FINE-GRAINED CONTEXTUAL COLLABORATIONS *w.r.t.* AN INPUT UNIT

In the above subsection, we introduce a method to distill knowledge of contextual collaborations into student nets of different regions. Given a student net, in this subsection, we develop an approach to disentangling from the student net explicit contextual collaborations *w.r.t.* a specific input unit u , *i.e.* identifying which input unit v collaborates with u to compute the network output.

We can consider a student net as a cascade of functions of N layers, *i.e.* $x^{(l)} = \phi_l(x^{(l-1)})$ (or $x^{(l)} = \phi_l(x^{(l-1)}) + x^{(l-m)}$ for skip connections), where $x^{(l)}$ denotes the output feature of the l -th layer. In particular, $x^{(0)}$ and $x^{(n)}$ indicate the input and output of the network, respectively. We only focus on a single scalar output of the network (we may handle different output dimensions separately if the network has a high-dimensional output). If the sigmoid/softmax layer is the last layer, we use the score before the softmax/sigmoid operation as $x^{(n)}$ to simplify the analysis.

3.2.1 PRELIMINARIES, THE ESTIMATION OF QUANTITATIVE CONTRIBUTION

As preliminaries of our algorithm, we extend the technique of (Shrikumar et al., 2016) to estimate the quantitative contribution of each neural activation in a feature map to the final prediction. We use $C_x \in \mathbb{R}^{H_l \times W_l \times D_l}$ to denote the contribution distribution of neural activations on the l -th layer $x \in \mathbb{R}^{H_l \times W_l \times D_l}$. The score of the i -th element C_{x_i} denotes the ratio of the unit x_i 's score contribution *w.r.t.* the entire network output score. Because $x^{(n)}$ is the scalar network output, it has a unit contribution $C_{x^{(n)}} = 1$. Then, we introduce how to back-propagate contributions to feature maps in low layers.

The method of contribution propagation is similar to network visualization based on gradient back-propagation (Mahendran & Vedaldi, 2015; Yosinski et al., 2015). However, contribution propagation reflects more objective distribution of numerical contributions over $\{x_i\}$, instead of biasedly boosting compacts of the most important activations.

Without loss of generality, in this paragraph, we use $o = \phi(x)$ to simplify the notation of the function of a certain layer. If the layer is a conv-layer or a fully-connected layer, then we can represent the convolution operation for computing each elementary activation score o_i of o in a vectorized form¹ $o_i = \sum_j x_j w_j + b$. We consider $x_j w_j$ as the numerical contribution of x_j to o_i . Thus, we can decompose the entire contribution of o_i , C_{o_i} , into elementary contributions of x_j , *i.e.* $C_{o_i \rightarrow x_j} = C_{o_i} \cdot \frac{x_j w_j}{o_i + \max\{-b, 0\}}$, which satisfies $C_{o_i \rightarrow x_j} \propto x_j w_j$ (see the appendix for details). Then, the entire contribution of x_j is computed as the sum of elementary contributions from all o_i in the above layer, *i.e.* $C_{x_j} = \sum_i C_{o_i \rightarrow x_j}$.

A cascade of a conv-layer and a batch-normalization layer can be rewritten in the form of a single conv-layer, where normalization parameters are absorbed into the conv-layer¹. For skip connections, a neural unit may receive contributions from different layers, $C_{x_j^{(l)}} = \sum_i C_{o_i^{(l+1)} \rightarrow x_j^{(l)}} + C_{x_j^{(l+m)}}$. If the layer is a ReLU layer or a Pooling layer, the contribution propagation has the same formulation as gradient back-propagations of those layers¹.

3.2.2 THE EXTRACTION OF CONTEXTUAL COLLABORATIONS

As discussed in (Bau et al., 2017), each neural activation o_i of a middle-layer feature o can be considered as the detection of a mid-level inference pattern. All input units must collaborate with neighboring units to activate some middle-layer feature units, in order to pass their information to the network output.

Therefore, in this research, we develop a method to

1. determine which mid-level patterns (or which neural activations o_i) the target unit u constitutes;
2. clarify which input units v help the target u to constitute the mid-level patterns;
3. measure the strength of the collaboration between u and v .

Let o^{bfr} and o denote the feature map of a certain conv-layer $o = f(x)$ when the network receives input features with the target unit u being activated and the feature map generated without u being activated, respectively. In this way, we can use $|o - o^{\text{bfr}}|$ to represent the absolute effect of u on the feature map o . The overall contribution of the i -th neural unit C_{o_i} depends on the activation score o_i , $C_{o_i} \propto \max\{o_i, 0\}$, where $\max\{o_i, 0\}$ measures the activation strength used for inference. The proportion of the contribution is affected by the target unit u can be roughly formulated as \tilde{C}_o .

$$\tilde{C}_{o_i} = \begin{cases} C_{o_i} \frac{|o_i - o_i^{\text{bfr}}|}{o_i}, & o_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $C_{o_i} = 0$ and thus $\tilde{C}_{o_i} = 0$ if $o_i \leq 0$, because negative activation scores of a conv-layer cannot pass information through the following ReLU layer (o is not the feature map of the last conv-layer before the network output).

In this way, \tilde{C}_{o_i} highlights a few mid-level patterns (neural activations) related to the target unit u . \tilde{C}_o measures the contribution proportion that is affected by the target unit u . We can use \tilde{C}_o to replace C_o and use techniques in Section 3.2.1 to propagate \tilde{C}_o back to input units $x^{(0)}$. Thus, $\tilde{C}_{x^{(0)}}$

¹Please see the Appendix for details.

represents a map of fine-grained contextual collaborations *w.r.t.* u . Each element in the map $\tilde{C}_{x_j^{(0)}}$ is given as $x_j^{(0)}$'s collaboration with u .

We can understand the proposed method as follows. The relative activation change $\frac{|o_i - o_i^{\text{bfr}}|}{o_i}$ can be used as a weight to evaluate the correlation between u and the i -th activation unit (inference pattern). In this way, we can extract input units that make great influences on u 's inference patterns, rather than affect all inference patterns. Note that both u and v may either increase or decrease the value of o_i . It means that the contextual unit v may either boost u 's effects on the inference pattern, or weaken u 's effects.

3.3 APPLICATION: EXPLAINING THE ALPHAGO ZERO MODEL

We use the ELF OpenGo (Tian et al., 2018; 2017) as the implementation of the alphaGo Zero model. We combine the above two methods to jointly explain each move's logic hidden in the value net of the alphaGo Zero model during the game. As we know, the alphaGo Zero model contains a value net, policy nets, and the module of the Monte-Carlo Tree Search (MCTS). Generally speaking, the superior performance of the alphaGo model greatly relies on the enumeration power of the policy net and the MCTS, but the value net provides the most direct information about *how the model evaluates the current state of the game*. Therefore, we explain the value net, rather than the policy net or the MCTS. In the ELF OpenGo implementation, the value net is a residual network with 20 residual blocks, each containing two conv-layers. We take the scalar output² before the final (sigmoid) layer as the target value to evaluate the current state on the Go board.

Given the current move of the game, our goal is to estimate unit-level contextual collaborations *w.r.t.* the current move. I.e. *we aim to analyze which neighboring stones and/or what global shapes help the current move make influences to the game*. We distill knowledge from the value net to student networks to approximate contextual collaborations within different regions. Then, we estimate unit-level contextual collaborations based on the student net.

Determining local contextual collaborations: We design two types of student networks, which receive lattices at the scales of 13×13 and 10×10 , respectively. In this way, we can conduct two distillation processes to learn neural networks that encode contextual collaborations at different scales.

As shown in Fig. 2, we have four student nets $\{f_i | i = 1, \dots, 4\}$ oriented to 13×13 lattices. Except for the output, the four student nets have the same network structure as the value net. The four student nets share parameters in all layers. The input of a student net only has two channels corresponding to maps of white stones and black stones, respectively, on the Go board. We crop four overlapping lattices at the four corners of the Go board for both training and testing. Note that we rotate the board state within each lattice I_{Λ_i} to make the top-left position corresponds to the corner of the board, before we input I_{Λ_i} to the student net. The neural network g has the same settings as the value net. g receives a concatenation of $[I_{\Lambda_1}, \dots, I_{\Lambda_4}]$ as the input. g outputs four scalar weights $\{\alpha_i\}$ for the four local student networks $\{y_i\}$. We learn g via knowledge distillation.

Student nets for 10×10 lattices have similar settings as those for 13×13 lattices. We divide the entire Go board into 3×3 overlapping 10×10 lattices. Nine student nets encode local knowledge from nine local lattices. We learn another neural network g , which uses a concatenation of $[I_{\Lambda_1}, \dots, I_{\Lambda_9}]$ to weight for the nine local lattices.

Finally, we select the most relevant 10×10 lattice and the most relevant 13×13 lattice, via $\max_i s_i$, for explanation.

Estimating unit-level contextual collaborations: In order to obtain fine-grained collaborations, we apply the method in Section 3.2.2 to explain two student nets corresponding to the two selected relevant lattices. We also use our method to explain the value net. We compute a map of contextual collaborations for each neural network and normalize values in the map. We sum up maps of the three networks together to obtain the final map of contextual collaborations \hat{C} .

²The value net uses the current state, as well as seven most recent states, to output eight values for the eight states. To simplify the algorithm, we take the value corresponding to the current state as the target value.

More specifically, given a neural network, we use the feature of each conv-layer to compute the initial \tilde{C}_o in Equation (4) and propagated \tilde{C}_o to obtain a map of collaborations $\tilde{C}_{x(o)}$. We sum up maps based on the 1st, 3rd, 5th, and 7th conv-layers to obtain the collaboration map of the network.

4 EXPERIMENTS

In experiments, we distilled knowledge of the value network to student nets, and disentangled fine-grained contextual collaborations *w.r.t.* each new move. We compared the extracted contextual collaborations and human explanations for the new move to evaluate the proposed method.

4.1 EVALUATION METRIC

In this section, we propose two metrics to evaluate the accuracy of the extracted contextual collaborations *w.r.t.* the new move. Note that *considering the high complexity of the Go game, there is no exact ground-truth explanation for contextual collaborations. Different Go players usually have different analysis of the same board state.* More crucially, as shown in competitions between the alphaGo and human players (alp, Retrieved 17 March 2016; 2017-05-27), the knowledge encoded in the alphaGo was sometimes beyond humans’ current understanding of the Go game and could not be explained by existing gaming principles.

In this study, we compared the similarity between the extracted contextual collaborations and humans’ analysis of the new move. The extracted contextual collaborations were just rough explanations from the perspective of the alphaGo. We expected these collaborations to be close to, but not exactly the same as human understanding. More specifically, we invited Go players who had obtained four-dan grading rank to label contextual collaborations. To simplify the metric, Go players were asked to label a relative strength value of the collaboration between each stone and the target move (stone), no matter whether the relationship between the two stones was collaborative or adversarial. Considering the double-blind policy, the paper will introduce the Go players if the paper is accepted.

Let Ω be a set of existing stones except for the target stone u on the Go board. $p_v \geq 0$ denotes the labeled collaboration strength between each stone $v \in \Omega$ and the target stone u . $q_v = |\hat{C}_v|$ is referred to as the collaboration strength estimated by our method, where \hat{C}_v denotes the final estimated collaboration value on the stone v . We normalized the collaboration strength, $\hat{p}_v = p_v / \sum_{v'} p_{v'}$, $\hat{q}_v = q_v / \sum_{v'} q_{v'}$ and computed the Jaccard similarity between the distribution of p and the distribution of q as the similarity metric.

In addition, considering the great complexity of the Go game, different Go players may annotate different contextual collaborations. Therefore, we also required Go players to provide a subjective rating for the extracted contextual collaborations of each board state, *i.e.* selecting one of the five ratings: 1-*Unacceptable*, 2-*Problematic*, 3-*Acceptable*, 4-*Good*, and 5-*Perfect*.

4.2 EXPERIMENTAL RESULTS AND ANALYSIS

Fig. 3 shows the significance of the extracted contextual collaborations, as well as possible explanations for contextual collaborations, where the significance of the stone v ’s contextual collaboration was reported as the absolute collaboration strength q_v instead of the original score \hat{C}_v in experiments. Without loss of generality, let us focus on the winning probability of the black. Considering the complexity of the Go game, there may be two cases of a positive (or negative) value of the collaboration score \hat{C}_v . The simplest case is that when a white stone had a negative value of \hat{C}_v , it means that the white stone decreased the winning probability of the black. However, sometimes a white stone had a positive \hat{C}_v . It may be because that this white stone did not sufficiently exhibit its power due to its contexts. Since the white and the white usually had a very similar number of stones in the Go board, putting a relatively ineffective white stone in a local region also wasted the opportunity of winning advantages in other regions in the zero-sum game. Similarly, the black stone may also have either a positive or a negative value of \hat{C}_v .

The Jaccard similarity between the extracted collaborations and the manually-annotated collaborations was 0.3633. Nevertheless, considering the great diversity of explaining the same game state,

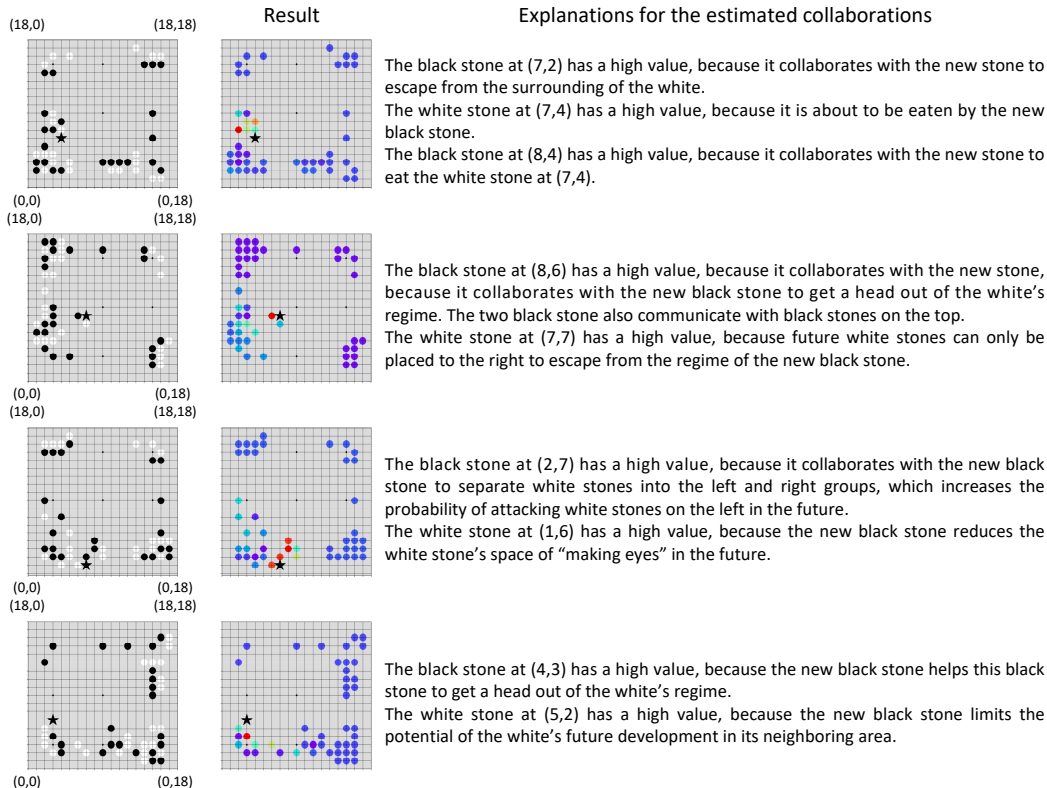


Figure 3: Significance of contextual collaborations *w.r.t.* the new black stone (the black star). Go players provided possible explanations for contextual collaborations. The red/blue color indicates a significant/insignificant contextual collaboration. Please see the appendix for more results.

the average rating score that was made by Go players for the extracted collaborations was 3.7 (between 3-*Acceptable* and 4-*Good*). Please see the appendix for more results.

5 CONCLUSION AND DISCUSSIONS

In this paper, we have proposed two typical methods for quantitative analysis of contextual collaborations *w.r.t.* a certain input unit in the decision-making of a neural network. Extracting fine-grained contextual collaborations to clarify the reason why and how an input unit passes its information to the network output is of significant values in specific applications, but it has not been well explored before, to the best of our knowledge. In particular, we have applied our methods to the alphaGo Zero model, in order to explain the potential logic hidden inside the model that is automatically learned via self-play without human annotations. Experiments have demonstrated the effectiveness of the proposed methods.

Note that there is no exact ground-truth for contextual collaborations of the Go game, and how to evaluate the quality of the extracted contextual collaborations is still an open problem. As a pioneering study, we do not require the explanation to be exactly fit human logics, because human logic is usually not the only correct explanations. Instead, we just aim to visualize contextual collaborations without manually pushing visualization results towards human-interpretable concepts. This is different from some previous studies of network visualization (Mahendran & Vedaldi, 2015; Yosinski et al., 2015) that added losses as the natural image prior, in order to obtain beautiful but biased visualization results. In the future, we will continue to cooperate with professional Go players to further refine the algorithm to visualize more accurate knowledge inside the alphaGo Zero model.

REFERENCES

- Alphago’s designers explore new ai after winning big in china. *In CADE METZ, BUSINESS*, 2017-05-27.
- Artificial intelligence: Google’s AlphaGo beats Go master Lee Se-dol*, Retrieved 17 March 2016.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *In CVPR*, 2017.
- Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *In NIPS*, 2016.
- Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. *In arXiv:1708.05493*, 2017.
- Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. *In CVPR*, 2016.
- Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *In arXiv:1704.03296v1*, 2017.
- Demis Hassabis and David Silver. Alphago zero: Learning from scratch. *In deepMind official website*, 18 October 2017.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β -vae: learning basic visual concepts with a constrained variational framework. *In ICLR*, 2017.
- Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric P. Xing. Harnessing deep neural networks with logic rules. *In arXiv:1603.06318v2*, 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 ICLR*, 2018.
- PangWei Koh and Percy Liang. Understanding black-box predictions via influence functions. *In ICML*, 2017.
- Devinder Kumar, Alexander Wong, and Graham W. Taylor. Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. *In CVPR Workshop on Explainable Computer Vision and Job Candidate Screening Competition*, 2017.
- Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. *In NIPS*, 2016.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *In NIPS*, 2017.
- Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *In CVPR*, 2015.
- Paulo E. Rauber, Samuel G. Fadel, Alexandre X. Falcão, and Alexandru C. Telea. Visualizing the hidden activity of artificial neural networks. *In Transactions on PAMI*, 23(1):101–110, 2016.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?” explaining the predictions of any classifier. *In KDD*, 2016.
- Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *In NIPS*, 2017.
- Ravid Schwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *In arXiv:1703.00810*, 2017.

- 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 ICCV*, 2017.
- Avanti Shrikumar, Peyton Greenside, Anna Y. Shcherbina, and Anshul Kundaje. Not just a black box: Interpretable deep learning by propagating activation differences. *in arXiv:1605.01713*, 2016.
- David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *In Nature*, 529(7587):484–489, 2016.
- Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. *In ICCV*, 2015.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *In arXiv:1312.6034*, 2013.
- S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. *In ECCV*, 2012.
- Austin Stone, Huayan Wang, Yi Liu, D. Scott Phoenix, and Dileep George. Teaching compositionality to cnns. *In CVPR*, 2017.
- Yuandong Tian, Qucheng Gong, Wenling Shang, Yuxin Wu, and C. Lawrence Zitnick. Elf: An extensive, lightweight and flexible research platform for real-time strategy games. *In Advances in Neural Information Processing Systems*, pp. 2656–2666, 2017.
- Yuandong Tian, Jerry Ma*, Qucheng Gong*, Shubho Sengupta, Zhuoyuan Chen, and C. Lawrence Zitnick. Elf opengo. <https://github.com/pytorch/ELF>, 2018.
- Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N. Nair. Explainable neural networks based on additive index models. *in arXiv:1806.01933*, 2018.
- Natalie Wolchover. New theory cracks open the black box of deep learning. *In Quanta Magazine*, 2017.
- Tianfu Wu, Xilai Li, Xi Song, Wei Sun, Liang Dong, and Bo Li. Interpretable r-cnn. *In arXiv:1711.05226*, 2017.
- Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *In ICML Deep Learning Workshop*, 2015.
- Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *In ECCV*, 2014.
- Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaro. Top-down neural attention by excitation backprop. *in ECCV*, 2016.
- Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. *In CVPR*, 2018a.
- Quanshi Zhang, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised learning of neural networks to explain neural networks. *in arXiv:1805.07468*, 2018b.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *In ICRL*, 2015.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *In CVPR*, 2016.

SUPPLEMENTARY MATERIALS FOR THE CONTRIBUTION PROPAGATION

Let $o = \omega \otimes x + \beta$ denote the convolutional operation of a conv-layer. We can rewrite this equation in a vectorized form as $\mathbf{o} = W\mathbf{x} + \mathbf{b}$, $\mathbf{o}, \mathbf{b} \in \mathbb{R}^{1 \times N}$, $W \in \mathbb{R}^{N \times N}$. For each output element o_i , $o_i = \sum_j x_j W_{ij} + b_i$. If the conv-layer is a fully-connected layer, then each element W_{ij} corresponds to an element in ω . Otherwise, W is a sparse matrix, *i.e.* $W_{ij} = 0$ if o_i and x_j are too far away to be covered by the convolutional filter.

Thus, we can write $o_i = \sum_j x_j w_j + b$ to simplify the notation. Intuitively, we can propagate the contribution of o_i to its compositional elements x_j based on their numerical scores. Note that we only consider the case of $o_i > 0$, because if $o_i \leq 0$, o_i cannot pass information through the ReLU layer, and we obtain $C_{o_i} = 0$ and thus $C_{o_i \rightarrow x_j} = 0$. In particular, when $b \geq 0$, all compositional scores just contribute an activation score $o_i - b$, thereby receiving a total contribution of $C_{o_i} \frac{o_i - b}{o_i}$. When $b < 0$, we believe the contribution of C_{o_i} all comes from elements of $\{x_j\}$, and each element's contribution is given a $C_{o_i} \cdot \frac{x_j w_j}{o_i - b}$. Thus, we get

$$C_{o_i \rightarrow x_j} = C_{o_i} \cdot \frac{x_j w_j}{o_i + \max\{-b, 0\}}$$

When a batch-normalization layer follows a conv-layer, then the function of the two cascaded layers can be written as

$$\begin{aligned} o_i &= \gamma \frac{(\sum_j x_j w_j + b) - \mu}{\sigma} + \beta' \\ &= \sum_j \left(\frac{\gamma w_j}{\sigma} \right) x_j + \left[\frac{\gamma(b - \mu)}{\sigma} + \beta' \right] \end{aligned}$$

Thus, we can absorb parameters for the batch normalization into the conv-layer, *i.e.* $w_j \leftarrow \frac{\gamma w_j}{\sigma}$ and $b \leftarrow \frac{\gamma(b - \mu)}{\sigma} + \beta'$.

For ReLU layers and Pooling layers, the formulation of the contribution propagation is identical to the formulation for the gradient back-propagation, because the gradient back-propagation and the contribution propagation both pass information to neural activations that are used during the forward propagation.

MORE RESULTS

Considering the great complexity of the Go game, there do not exist ground-truth annotations for the significance of contextual collaborations. Different Go players may have different understanding of the same Go board state, thereby annotating different heat maps for the significance of contextual collaborations. More crucially, our results reflect the logic of the automatically-learned alphaGo Zero model, rather than the logic of humans.

Therefore, in addition to manual annotations of collaboration significance, we also require Go players to provide a subjective evaluation for the extracted contextual collaborations.

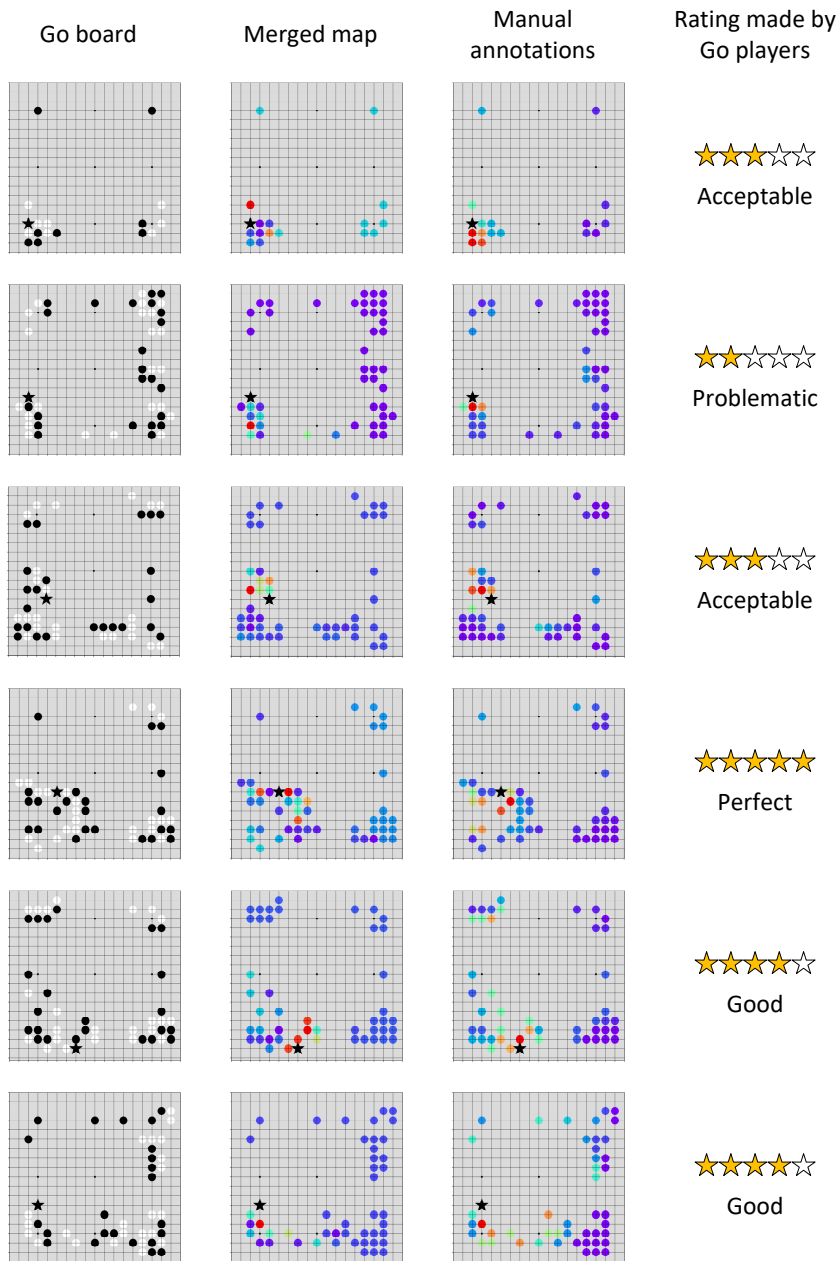


Figure 4: We compared the extracted contextual collaborations at different scales (the second, third, fourth, and fifth columns) with annotations made by Go players.

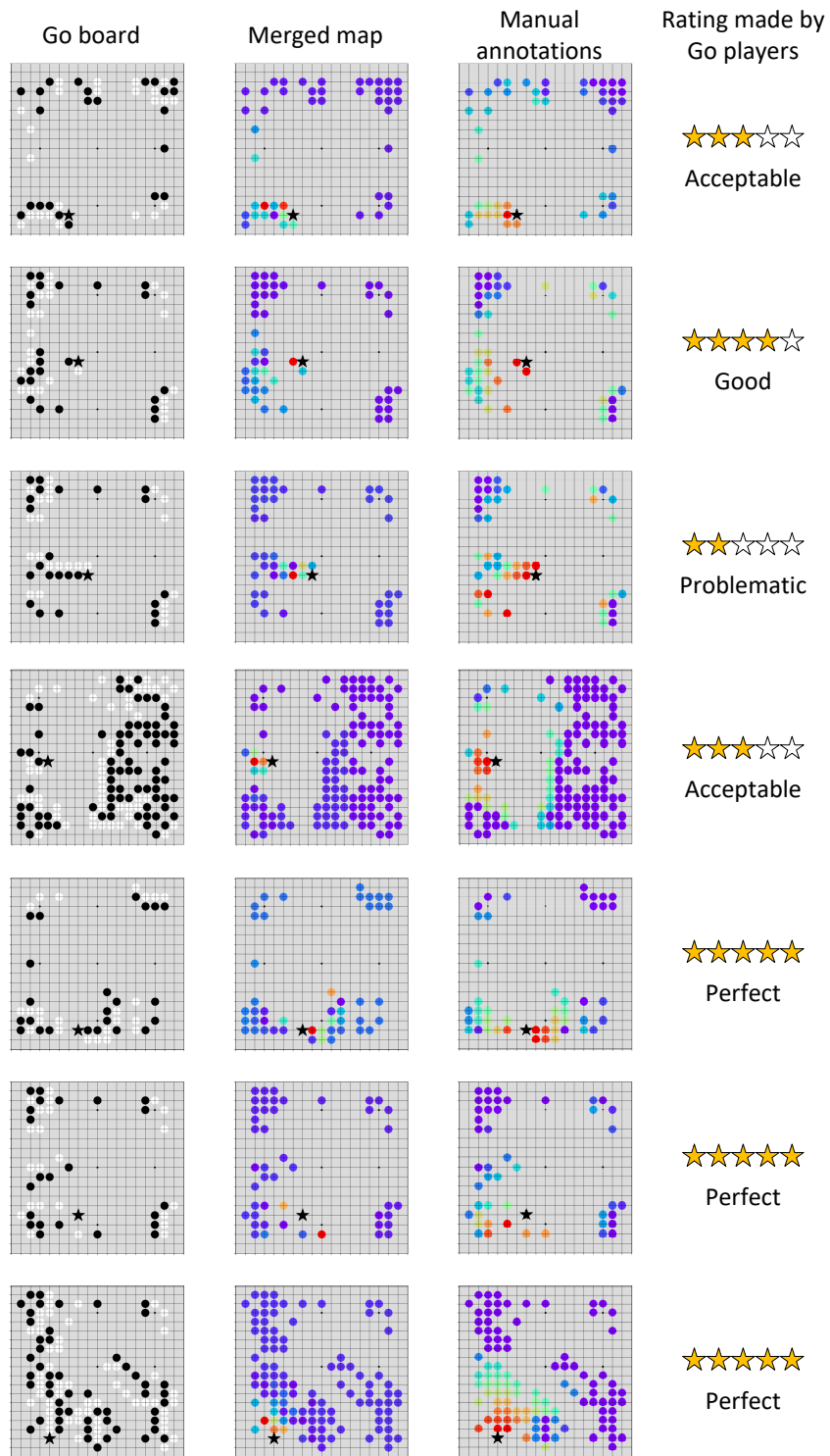


Figure 5: We compared the extracted contextual collaborations at different scales (the second, third, fourth, and fifth columns) with annotations made by Go players.

CONTEXTUAL COLLABORATIONS OF LOCAL REGIONS

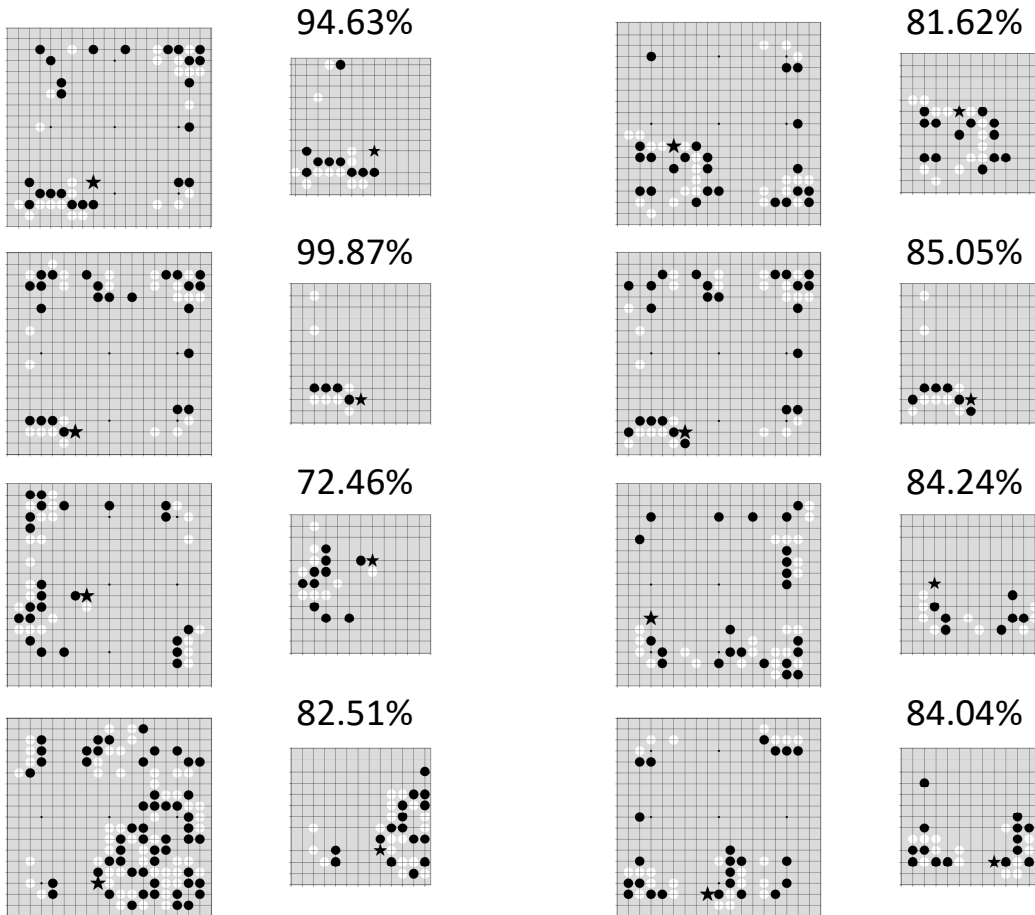


Figure 6: We show the significance of contextual collaborations within a local lattice. The score for the i -th lattice is reported as $\frac{s_i}{\sum_j s_j}$.