Interpretable Convolutional Neural Network Including Attribute Estimation for Image Classification

Kazaha Horii †, Keisuke Maeda ††, Takahiro Ogawa (member) ††, Miki Haseyama (member) ††

Abstract An interpretable convolutional neural network (CNN) including attribute estimation for image classification is presented in this paper. Although CNNs perform highly accurate image classification, the reason for the classification results obtained by the neural networks is not clear. In order to provide interpretation of CNNs, the proposed method estimates attributes, which explain elements of objects, in an intermediate layer of the network. This enables improvement of the interpretability of CNNs, and it is the main contribution of this paper. Furthermore, the proposed method uses the estimated attributes for image classification in order to enhance its accuracy. Consequently, the proposed method not only provides interpretation of CNNs but also realizes improvement in the performance of image classification.

Key words: Interpretable convolutional neural network, attribute estimation, image classification.

1. Introduction

There is now a huge number of images on the Web due to the development of imaging devices and spread of social networking services. In order to efficiently browse desired images, image retrieval technologies are necessary. Class information representing objects in images is often used for image retrieval. Since it is necessary to give classes to images in advance for image retrieval, studies on automatic image classification by classes have recently been conducted. Convolutional neural networks (CNNs) are the most widely used deep learning methods, and highly accurate image classification is realized by using CNNs. CNNs can transform visual features to effective features for image classification by constructing a large number of hidden layers. The lower convolution layers capture basic features (e.g., colors and shapes), and the top layer can represent more complicated structures. Furthermore, it has been reported that the learned features obtained from these networks were suitable for many other problems. However, there is no clear understanding of the reasons for the classification results obtained by deep learning-based neural networks. For example, CNNs are end-to-end systems, that is, the meaningful information in the networks is only input images and output classes. In other words, there is still little insight into the internal operation and behavior for human interpretation of CNNs. This problem also causes a limitation of improvement of CNNs’ classification performance since CNNs can be trained by only increasing the number of pairs, i.e., input images and output classes. Therefore, it is necessary to provide an interpretation of CNNs by using alternative approaches.

To provide explanations for the deep neural networks, many interesting approaches have been proposed. They focused on feature visualization that can answer questions about what the CNN is looking for. Kindermans et al. proposed “PatternAttribution” which performed visualization using a heat map of pixel-wise reliability of classification results by CNNs. Springenberg et al. proposed a deconvolution approach to visualize concepts learned by neurons in higher of a CNN. “Grad-cam” and “Smooth-Grad” are gradient-based visualization approaches that highlight the important regions for image classification. “LIME” is a method to visually explore the behavior of the models from variously deformed input. These conventional methods focus on visualization of what the CNN has learned. They can efficiently pro-
provide interpretability, and humans can understand the results of the CNN’s outputs. On the other hand, more specific and understandable information such as textual information would be useful to find the reason for the deep neural network-based classification results.

Generally, objects are identified on the basis of high-level descriptions called attributes that explain their elements. Since attributes correspond to high-level properties of the objects that can be understood by both machines and humans, attributes have attracted much attention in visual recognition research due to the fact that the attributes of objects provide detailed knowledge about the objects. Thus, attributes can improve not only object recognition performance of machines but also the interpretability of the classification results. In fact, attributes have been used for describing images for retrieval, and for classification. For provision of interpretation, although visualization methods such as those in references provide what machines see to humans, these visualization methods have ambiguity for human interpretation of CNNs. On the other hand, attributes have no ambiguity due to the verbal description of objects. Therefore, attributes are important for understanding object appearance that provides interpretation of CNNs without ambiguity. Moreover, the performance with multi-task learning, which constructed CNNs based on both multiple information of classes and information related to these classes, was higher than that of single-task learning. Thus, multi-task CNNs for image classification and attribute estimation improve the interpretability and performance of CNNs. Specifically, since attributes have potentials to bridge low-level features, that is, visual features and the highest-level class information, attributes can be estimated by the features of an intermediate layer between input and class output layers.

In this paper, we propose a novel interpretable CNN including attribute estimation for image classification. We estimate attributes in an intermediate layer of a CNN to provide interpretable results by focusing on the attribute potentials to bridge between visual features and class information. This is the biggest contribution of this paper. Moreover, in order to enhance classification performance of object classes, the proposed method uses the estimated attributes as features in addition to the visual features in the CNN architecture. Consequently, the proposed method can not only clarify the reason for the classification results by estimating the attributes but also realize more accurate image classification by using both the estimated attributes and the visual features. Actually, as shown in the experiment, the performance of the image classification can be improved by about 26% compared to that of image classification using only visual features.

In order to reveal our position, we show a research map in Fig. 1. Although various conventional methods focus on feature visualization for interpretation of CNNs, we focus on text-based description. The proposed method shows not only classification results but also estimated attributes for interpretation of CNNs as a text-based description, that is, the proposed method tackles a multi-label task. Note that since a single-label task which is equal to general image classification can output only classification results, it is impossible to apply a method to give text-based descriptions for interpretation of CNNs. Unlike conventional multi-task learning, the proposed method uses the results obtained from one task (attribute estimation) for other tasks (image classification). This is the novelty of the proposed method.

In the previous conference paper we reported, we focused on the attribute estimation and performed the experiment about only attribute estimation. In this paper, we focus on the entire framework, that is, we perform not only attribute estimation but also final image classification. It is expected that the proposed method realizes both the improvement of image classification performance and provision of interpretability.

2. Interpretable CNN for Image Classification

In this section, we explain the novel interpretable CNN including attribute estimation for image classification. As shown in Fig. 2, the proposed method estimates attributes as the intermediate layers’ outputs and uses the estimated attributes with the visual features for image classification. Specifically, the proposed method estimates attributes via fine-tuning of a pre-trained CNN. Although general CNNs focus on a single label problem, since one image has multiple attributes, we construct CNNs for a multi-label problem. Then, by using both the estimated attributes and the visual features, the proposed method performs image classification. Since class information is based on visual features and attributes, it is considered that correlation between visual features and attributes is latent high. Therefore, by maximizing canonical correlation between two
set of features, it is realized to calculate discriminative features for image classification. Moreover, Yeh et al. reported that canonical features obtained via canonical correlation analysis (CCA)\(^\text{20}\) between two sets of features which have relevancy are effective for image classification\(^\text{20}\). Therefore, in order to enhance the performance of image classification, the proposed method calculates the canonical features between the visual features obtained from the pre-trained CNN and the estimated attributes. The proposed method inputs the visual features and the canonical features obtained by projecting the attributes into a single feed-forward network-based classifier.

2.1 Attribute Estimation

The proposed method estimates attributes by performing fine-tuning of the pre-trained CNN. Specifically, the proposed method adds full-connected layers for multi-label classification to an intermediate layer of the pre-trained CNN and performs the training by fine-tuning of the full-connected layers. In the training phase, we construct the network for multi-label classification based on the reference\(^\text{20}\). Specifically, we use a sigmoid function as an activation function in order to output a probability value \(\mathbf{\alpha}_n^{(a)}\) from the full-connected layers as follows:

\[
\mathbf{\alpha}_n^{(a)} = \text{sigmoid}(\mathbf{x}_n) = \frac{1}{1 + \exp(-\mathbf{x}_n)},
\]  

where \(\mathbf{\alpha}_n^{(a)} \in \mathbb{R}^{C_a}\) (\(C_a\) being the number of attributes), and \(\mathbf{x}_n \in \mathbb{R}^d\) \((n = 1, 2, \cdots, N; N\) being the number of training images) is a feature vector obtained from the intermediate layer of the pre-trained CNN. Note that \(\mathbf{x}_n\) is standardized. We train the network based on a cross-entropy loss\(^\text{21}\) between the output \(\mathbf{\alpha}_n^{(a)}\) and the ground truth \(\mathbf{y}_n^{(a)} \in \mathbb{R}^{C_a}\), where \(\mathbf{y}_n^{(a)}\) is a binary vector. In the test phase, given a visual feature vector \(\mathbf{x}\) from the intermediate layer, we calculate the probability value \(\mathbf{\alpha}_n^{(a)} = [\mathbf{\alpha}_1^{(a)}, \mathbf{\alpha}_2^{(a)}, \cdots, \mathbf{\alpha}_{C_a}^{(a)}]\) in the same manner as Eq. (1). Finally, we estimate the attributes \(\mathbf{o}_n^{(a)} = [o_1^{(a)}, o_2^{(a)}, \cdots, o_{C_a}^{(a)}]\) by the following equation:

\[
o_n^{(a)} = \begin{cases} 1 & \mathbf{\alpha}_n^{(a)} \geq Th \\ 0 & \mathbf{\alpha}_n^{(a)} < Th \end{cases},
\]

where \(c = 1, 2, \cdots, C_a\) and \(Th\) is a threshold value. If \(o_n^{(a)} = 1\), it is assumed that the target image includes the attribute corresponding to \(c\).

The proposed method constructs the network for attribute estimation by using the intermediate layer of the CNN. By adding the full-connected layer for the multi-label image classification and performing the fine-tuning, multiple attribute estimation is realized. Therefore, attributes that represent elements of the image enable interpretability of the CNN. The details are discussed in the following subsection.

2.2 Image Classification from Visual Features and Estimated Attributes

In this subsection, we explain the image classification from visual features and estimated attributes. The proposed method uses both the visual features and attributes and transforms the attributes to canonical features via the CCA’s projection considering the relationships between these two features. Furthermore, the proposed method constructs a classifier based on a single feed-forward network\(^\text{22} \text{24}\) for image classification.

In the training phase, given a visual feature vector \(\mathbf{x}_n\) from the intermediate layer of the pre-trained CNN and the ground truth attribute vector \(\mathbf{y}_n^{(a)}\), a canonical feature vector \(\mathbf{\hat{y}}_n^{(a)}\) between the visual features and the
attributes is calculated. For calculating the canonical feature vector, projections $p_x$ and $p_y$ are estimated by maximizing the canonical correlation between the visual features and the attributes as follows:

$$
(p_x, p_y) = \arg \max_{p_x,p_y} \frac{p_x^T C_{xy} p_y}{\sqrt{p_x^T C_{xx} p_x} \sqrt{p_y^T C_{yy} p_y}}, \quad (3)
$$

where $C_{xx} = XX^\top$, $C_{yy} = YY^a$, $C_{xy} = XY^a$. A visual feature matrix $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d_x \times N}$ is defined by aligning the visual feature vectors $x_n$, and $Y^a = [y_1^{(a)}, y_2^{(a)}, \ldots, y_N^{(a)}] \in \mathbb{R}^{a \times N}$ ($a$ being the number of attributes) is defined in the same manner. The optimization problem in Eq. (3) can be solved as a generalized eigenvalue problem\(^{33}\). Then we define a projection matrix $P_y = [p_{y,1}, p_{y,2}, \ldots, p_{y,d_y}] \in \mathbb{R}^{a \times d_y}$ ($d_y$ being the dimension of a canonical feature vector) by aligning the eigenvectors for which the corresponding eigenvalues are higher than the others. We calculate the canonical features by $\hat{y}_n^{(a)} = P_y^T y_n^{(a)}$ and define a new concatenated feature vector $z_n = [x_n^\top, \hat{y}_n^{(a)}]^\top \in \mathbb{R}^{d_z}$, where $d_z = d_x + d_y$.

We input $z_n$ into the single feed-forward network classifier for training the network based on the reference\(^{34}\). The aim of training the classifier is calculation of a weight matrix $\beta \in \mathbb{R}^{C \times L_c}$ ($L_c$ being the number of nodes of the hidden layer and $C$ being the number of classes) between the hidden layer and the output layer. In order to calculate the weight matrix $\beta$, we try to minimize the training error between $o_n^{(c)} = [o_{n,1}^{(c)}, o_{n,2}^{(c)}, \ldots, o_{n,C}^{(c)}]^\top$ and the ground truth $y_n^{(c)} \in \mathbb{R}^C$ as well as the output weights as follows:

$$
\arg \min_{\beta} \frac{1}{2} \|\beta\|^2_F + \frac{\mu}{2} \sum_{n=1}^N W_{nn} \|o_n^{(c)}\|^2 \\
\text{s.t. } \beta h(z_n) = y_n^{(c)} - o_n^{(c)}, \quad (4)
$$

where $\mu$ is a regularization parameter. Since the number of training images is different for each class, we adopt a weight diagonal matrix $W = diag(W_{nn})$ for solving the class-imbalanced problem. In order to deal with the classes that have a small amount of data, we define the diagonal elements of the weight matrix $W$ according to the reference\(^{34}\) as follows:

$$
W_{nn} = \frac{1}{N_c}, \quad (5)
$$

where $N_c$ is the number of images of the belonging class. This provides optimized $\beta$ and is expected to prevent accuracy degradation due to the class imbalance problem. In case that the diagonal matrix $W$ is an identity matrix, the classifier is the same as that of reference\(^{32}\), and it is not able to consider the class imbalance problem. Furthermore, $h(z_n) \in \mathbb{R}^{L_c}$ is an output vector of the hidden layer, which is defined as

\[\text{Interpretable CNN by attribute estimation} \quad \text{Full-connected layers for multi-label classification} \quad \text{Projection obtained via CCA between visual features and attributes} \quad \text{Use of estimated attributes for improving classification performance} \quad \text{Fig. 2 Architecture of the proposed network. Firstly, the proposed method estimates attributes by applying fine-tuning to full-connected layers connected to the intermediate layer of the pre-trained CNN for multi-label image classification. Secondly, the proposed method calculates a projection $P_y$ by applying CCA to the visual features and estimated attributes. Then canonical features are calculated by using the projection $P_y$ and the estimated attributes. Finally, the proposed method concatenates the visual features and the canonical features and inputs these features into a single feed-forward network-based classifier.}\]
\[ h(z_n) = [h_1(z_n), h_2(z_n), \ldots, h_{L-1}(z_n)]^T. \]  

(6)

Note that \( h(z_n) \) actually maps the features from the \( d_z \)-dimensional input space to the \( L \)-dimensional hidden feature space by using the Gaussian function. The optimal solution of \( \hat{\beta} \) in Eq. (4) can be obtained as

\[ \hat{\beta} = Y(c)W\frac{I}{\mu} + H^THW)^{-1}H^T, \]

(7)

where \( H = [h(z_1), h(z_2), ..., h(z_N)] \).

Given a new test visual feature vector \( x \) and its estimated attributes \( o^{(a)} \) obtained as described in the previous subsection, we calculate a canonical feature vector as \( \hat{\theta}^{(a)} = P_y^{(a)} o^{(a)} \) and define a concatenated feature vector \( z = [x^T, \hat{\theta}^{(a)}]^T \). Then, by using the \( \hat{\beta} \) obtained in Eq. (7), the output vector \( o^{(c)} = [o_1^{(c)}, o_2^{(c)}, \ldots, o_C^{(c)}]^T \) is obtained as

\[ o^{(c)} = \hat{\beta}h(z). \]

(8)

The final classification result, i.e., the class label of \( x \), is obtained as

\[ \text{class} = \arg \max_c o^{(c)}. \]

(9)

As described in the previous subsection and this subsection, the proposed method estimates the class labels with their corresponding attributes. Since the attributes represent elements of the image and bridge the visual features and the class information, using this information is effective for improving the interpretability of image classification results. Furthermore, by using the canonical features obtained via CCA between the visual features and the attributes, the proposed method enables more accurate image classification than do methods using only visual features.

3. Experimental Conditions

In order to verify the effectiveness of the proposed method, we used Clothing Attributes Dataset\(^{36}\), which contains 1856 images and 26 labels. The ground truth was provided on image-level. The dataset was annotated by six persons. When all annotators did not agree on their annotations, they gave “Nan” for the label of the image. We used labels about “Category” as classes (seven classes) and others (25 labels) as attributes. The numbers of images for the classes are shown in Table 1, and the labels used as attributes are shown in Table 2. Note that we removed some images with “Nan Category” and obtained 1104 images from the dataset. Most of the attributes excluding the attributes shown in Table 3 have “Yes”/“No”/“Nan”. Thus, each attribute other than those listed in Table 3 has three dimensions whose elements are 1 or 0. For example, for “Black”, if an image has a black color, the image has “Yes Black”. If an image does not have a black color, the image has “No Black”. If an image is undecidable, the image has “Nan Black”. Therefore, the dataset had 77 attributes. Since interpretability may vary by individual users, it is difficult to determine uniquely. Therefore, in this experiment we use “Yes/No/Nan” in order to consider this point and flexibly deal with the case where it is difficult to determine attributes uniquely. We randomly selected 964 images for training images and 140 images for test images. As a validation dataset, we randomly selected ten images from the training images for each class.

In the proposed method, we used a pre-trained Inception-v3\(^{37}\) model based on ImageNet\(^2\). A CNN pre-trained by ImageNet is considered to be able to fully express the images used in this experiment. Moreover, since the proposed method performs both attribute estimation and image classification, it may cause overfitting and decrease of attribute estimation performance by fine-tuning the CNN based on only classes. Therefore, we used the pre-trained CNN as a feature extractor. We added a full-connected layer to the third pooling layer of Inception-v3. In this experiment, we determined parameters in such a way that the proposed method output the best classification performance for the validation dataset. The parameters in comparative methods were set in the same manner. We searched the optimal parameters in the ranges shown in Table 4. We evaluated the performance of the final class estimation in the proposed method by using F-measure, which is a harmonic mean of Recall and Precision used for image classification. For attribute estimation, we evaluated the performance of the proposed method by using evaluation indexes following the reference\(^{38}\). Specifically, Accuracy, Macro-P, Macro-R, Macro-F, Micro-F and Hamming loss, which are generally used for multi-label classification problems, were adopted in the experiment.

We used some comparative methods (CMs) in the experiment. CM1 uses features obtained by the concatenation of the visual features \( x \) and the estimated attributes \( o^{(a)} \). CM1 inputs \( \tilde{x} = [x^T, o^{(a)}]^T \) into the classifier. Although CM1 estimates attributes and performs image classification by using visual features and estimated attributes like the proposed method, CM1 does not use CCA before inputting the two sets of features into the classifier. CM2 uses only visual
The proposed method and the comparative methods were trained by using a personal computer (GPU) with Intel (R) Core (TM) i7-4790K CPU@4.00GHz with 512 Gbytes of RAM and GPU GeForce GTX 1080 Ti.

4. Performance Evaluation

Table 5 shows classification results obtained by using ground truth attributes (GT-Att); that is, the results of the third and second columns show the ideal and upper limits of the proposed method (PM) and CM1. It is assumed that the attribute estimation is performed completely. Specifically, “VF”, “VF and GT-Att” and “VF and canonical features” are equivalent to CM2, CM1 with ideal attribute estimation and PM with ideal attribute estimation, respectively. When the visual features and the canonical features obtained by using the ground truth attributes are given, F-measure can be improved by about 26% compared to that using only visual features. Thus, it is verified that the use of combination of visual features and attributes is effective.

Table 6 shows the final classification results of the proposed method and the comparative methods. The threshold value parameter was set to a value which provided the best performance for each method in the test phase. Since PM is superior to CM5 which is a conventional multi-task learning method of image classification, the use of the estimated attributes realizes more accurate image classification than non-use of the estimated attributes. Furthermore, since PM is superior to CM4, the proposed method is more effective than the method which simultaneously treats classes and attributes as the multi-label problem. Since PM is superior to CM2 and CM3, the combination use of visual features and the estimated attributes realizes more accurate image classification than that using only visual features or the estimated attributes. Therefore, the use of the estimated attributes improves the image classification performance. Moreover, Fig. 3 shows examples that are classified correctly by PM but incorrectly by CM1. Even if some attributes are incorrectly estimated, PM correctly classifies images but CM1 incorrectly clas-
sifies images. It is verified that applying CCA to visual features and the estimated attributes is effective for the following image classification.

Furthermore, by outputting estimated attributes, we can understand what information is considered in the intermediate layer of the CNN. Figure 4 shows examples that are classified correctly by PM but incorrectly by CM2. PM not only improves the image classification performance but also shows the estimated attributes. This means that PM realizes more accurate image classification than that by using only visual features and can also provide interpretability of the CNN. Furthermore, we conducted a subject experiment for measuring the interpretability of attributes estimated by PM and CM5. Seven subjects participated, and we showed 12 images with classification results and estimated attributes. Note that in order to prevent effects on classification results, we randomly selected images that were correctly or incorrectly classified by both PM and CM5. We asked for subjects which method had better interpretability. Note that we hid the information about which method was PM or CM5 when this experiment was conducted. The average of percentage of subjects who judge that PM is more interpretable than CM5 is 61.3%. Therefore, PM tends to be more effective than CM5.

This paragraph discusses confusion matrices of image classification results as shown in Fig. 5. The image classification results of CM3 and CM4 are worse than other methods. It is considered that since the performance of the attribute estimation is not enough, and CM3 classifies images based only on the estimated attributes, CM3 is directly influenced by the attribute estimation performance. CM4 that tackles a multi-label task with vector concatenation of classes and attributes is directly influenced by the class imbalance problem. Although the image classification performance of PM, CM1, CM2 and CM5 are similar, it can be confirmed that PM, CM1 and CM2 that use a classifier based on the reference\(^\text{13}\) are able to consider the class imbalance slightly compared to CM5.

This paragraph discusses the relationship between the results of final classification and those of attribute estimation. Figure 6 shows the results of the proposed method and comparative methods that were obtained by changing the threshold value (\(Th\)). The figure shows the relationships between the threshold value (\(Th\)) and the final classification performance. Since there is no threshold parameter in CM2 and CM4, the F-measures are constant. Figures 7 and 8 show the results of attribute estimation. Note that since the methods used for attribute estimation in PM, CM1 and CM3 are the same, the results of attribute estimation are the same. The F-measure of the image classification results changes drastically in Fig. 6 according to the threshold values. When the threshold value changes from 0.7 to 0.8, all of the results of attribute estimation and F-measure of image classification become worse. The image classification performance and the results of attribute estimation are related.

This paragraph discusses the results of attribute estimation. Figure 9 shows some of the attributes estimated to be given by PM with \(Th = 0.3\). Note that this figure shows examples of the estimation results for some attributes and classes. Attributes that are commonly given such as “No Floral” and “V-shapes” are estimated to be given commonly in every class. Class-specific attributes such as “No Collar” and “Necktie” are estimated differently for each class. The proposed method sets the same threshold value for all attributes. More accurate attribute estimation will be realized by adaptively setting the threshold values for each attribute, and more accurate image classification will also be realized.

The proposed method has some limitations.

- First, only a full-connected layer was added to the proposed method and only the added layer was fine-tuned for attribute estimation. Since pre-trained CNNs are trained on the basis of the class information, it is necessary to transform the intermediate layer’s features to features suitable for attribute estimation. Thus, the use of some hidden layers trained for attribute estimation will enable the performance to be enhanced.
- Second, since there is a large number of attributes, it is difficult to prepare a dataset with attributes that are sufficient for the explanation of objects. Thus, we need to introduce a method such as zero-shot learning\(^{13}\) or weakly supervised learning\(^{42}\) for attribute estimation in order to tackle that problem.
- Third, the proposed method might not use optimal pre-trained CNNs for the dataset. In references\(^{40}\) it was reported that pre-trained CNNs should be adaptively selected for application domains, and improvement of attribute estimation and image classification performance can be realized. The performance of the proposed method can be improved by constructing application-specific pre-trained CNNs.

Finally, we show our future works. As mentioned
Table 5 Classification results (F-measures) with the proposed classifier using visual features (VF) and ground truth attributes (GT-Att). “VF”, “VF and GT-Att” and “VF and canonical features projected GT-Att” are equivalent to CM2, CM1 with ideal attribute estimation and PM with ideal attribute estimation, respectively.

|                | VF           | VF and GT-Att | VF and canonical features projected GT-Att |
|----------------|--------------|---------------|------------------------------------------|
|                | (CM2)        | (Ideal of CM1)| (Ideal of PM)                            |
|                | 0.429        | 0.479         | 0.543                                    |

Table 6 Classification results (F-measures) of the proposed method (PM) and the comparative methods (CMs). A parameter of threshold value \((Th)\) was set to a value which provided the best performance for each method in the test phase.

|     | PM        | CM1       | CM2       | CM3       | CM4       | CM5       |
|-----|-----------|-----------|-----------|-----------|-----------|-----------|
|     | 0.486     | 0.461     | 0.429     | 0.274     | 0.257     | 0.471     |

(a) Sweater ✓ “No Black”, “No Cyan”, ...
✗ “No gray”, “No Placket”, ...

(b) Sweater ✓ “Black”, “No Necktie”, ...
✗ “Solid”, “Placket”, ...

(c) Shirt ✓ “No Necktie”, “Collar”, ...
✗ “No Stripe”, “Low Exposure”, ...

(d) Dress ✓ “No Collar”, “Female”, ...
✗ “No Solid”, “Placket”, ...

Fig. 3 Examples that are classified correctly by PM but incorrectly by CM1. ✓ is an estimated attribute correctly, and ✗ is an estimated attribute incorrectly. Some of the representative attributes estimated by the proposed method are shown.

above, we will introduce methods of feature transformation for attribute estimation such as constructing some hidden layers. There is room for improvement of attribute estimation performance. There is also room for improvement of image classification performance. There are various kinds of CCA such as Multiset CCA\(^{47}\), Discriminate Locality Preserving CCA\(^{48}\), Supervised Locality Preserving CCA\(^{49}\) and Supervised
Multi View CCA\textsuperscript{50}. We will introduce extended versions of the CCA for image classification. In this experiment, we used binarized attribute outputs for image classification in order to show the estimated attributes clearly. It is considered that the proposed method with binarized attribute outputs is the simplest for improvement of interpretability. However, it is not always necessary to use binarized attribute outputs for final image classification. Although the experimental results in this paper verify the improvement of both image classification performance and interpretability by using the binarized attribute outputs, further improvement of the proposed method is expected by using the continuous value outputs. Furthermore, we will use other kinds of images in order to confirm the generalization performance of the proposed method. Moreover, improvement of image classification and attribute estimation performance is expected by incorporating processing specific to a target dataset like the previous methods\textsuperscript{36,39,41} into the proposed method.

5. Conclusions

In this paper, we have proposed an interpretable image classification with attribute information. The proposed method estimates attributes in the intermediate layer and uses the estimated attributes for image classification. The main contribution of this paper is the attribute estimation in the intermediate layer, and we can provide interpretation of CNNs for humans. Furthermore, by using the estimated attributes, the proposed method realizes more accurate image classification than that by using only visual features.
6. Acknowledgment

Funding: This work was partly supported by Japan Society for the Promotion of Science (grant number JP17H01744 and JP18J10373), MIC/SCOPE (grant number #181601001 and #181503004) and the Photo-excitonix Project in Hokkaido University.

References

1) J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
2) A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.
3) A. Abdulkadi, G. Wang, J. Lu, and K. Jia, “Multi-task CNN model for attribute prediction,” IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1949–1959, 2015.
4) M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717–1724.
5) A. Sharif, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 806–813.
6) Z. Lipton, “The mythos of model interpretability,” arXiv preprint arXiv:1606.03490, 2016.
7) M. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proceedings of European Conference on Computer Vision, 2014, pp. 818–833.
8) P. Kindermans, K. Schütt, M. Alber, K. Erhan, B. Kim, and S. Dähne, “Learning how to explain neural networks: Patternnet and patternattribute,” arXiv preprint arXiv:1705.05598, 2017.
9) M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.
10) R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 618–626.
11) D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” arXiv preprint arXiv:1706.05885, 2017.
12) J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” arXiv preprint arXiv:1412.6806, 2014.
13) C. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 453–465, 2014.
14) Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, “Label-embedding for attribute-based classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 819–826.
15) A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1778–1785.
16) C. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.
17) V. Ferrari and A. Zisserman, “Learning visual attributes,” in Advances in Neural Information Processing Systems, 2008, pp. 433–440.
18) M. Douze, A. Ramisa, and C. Schmid, “Combining attributes and fisher vectors for efficient image retrieval,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 745–752.
19) N. Kumar, P. Belhumeur, and S. Nayar, “Facetracer: A search engine for large collections of images with faces,” in Proceedings of European Conference on Computer Vision, 2008, pp. 340–353.
20) B. Sridhique, R. Feris, and L. Davis, “Image ranking and retrieval based on multi-attribute queries,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 801–808.
21) D. Mahajan, S. Sellamanickam, and V. Nair, “A joint learning framework for attribute models and object descriptions,” in Proceedings of International Conference on Computer Vision, 2011, pp. 1227–1234.
22) G. Wang and D. Forsyth, “Joint learning of visual attributes, object classes and visual saliency,” in Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 537–544.
23) Y. Wang and G. Mori, “A discriminative latent model of object classes and attributes,” in Proceedings of European Conference on Computer Vision. Springer, 2010, pp. 155–168.
24) S. Ruder, “An overview of multi-task learning in deep neural networks,” arXiv preprint arXiv:1706.05998, 2017.
25) T. Takada, N. Inoue, T. Ya, Y. Hasegaki, and K. Aizawa, “Similar floor plan retrieval featuring multi-task learning of layout type classification and room presence prediction,” in Proceedings of IEEE International Conference on Consumer Electronics, 2018, pp. 1–6.
26) J. Chen and C. Ngo, “Deep-based ingredient recognition for cooking recipe retrieval,” in Proceedings of the 26th ACM International Conference on Multimedia, 2016, pp. 32–41.
27) K. Horii, K. Maeda, T. Ogawa, and M. Haseyama, “A note on attribute estimation for improving interpretability of convolutional neural network (in japanese),” The Institute of Image Information and Television Engineers Technical Report, pp. 275–279, 2019.
28) H. Hotelling, “Relations between two sets of variates,” Biometrika, vol. 28, no. 3/4, pp. 321–377, 1936.
29) Y. Yeh, C. Huang, and Y. Wang, “Heterogeneous domain adaptation and classification by exploiting the correlation subspace,” IEEE Transactions on Image Processing, vol. 23, no. 5, pp. 2009–2018, 2014.
30) R. Bartyzel, “Multi-label image classification with Inception net,” https://towardsdatascience.com/multi-label-image-classification-with-inception-net-cbb2ee538e30, 2017 (accessed 1 March 2018).
31) P. De Boer, D. Kroese, S. Mannor, and R. Rubinstein, “A tutorial on the cross-entropy method,” Annals of Operations Research, vol. 134, no. 1, pp. 19–67, 2005.
32) G. Huang, Q. Zhu, and C. Siew, “Extreme learning machine: A new learning scheme of feedforward neural networks,” in Proceedings of IEEE International Joint Conference on Neural Networks, 2004, vol. 2, pp. 985–990.
33) W. Schmidt, M. Kraaijveld, and R. Duin, “Feedforward neural networks with random weights,” in Proceedings of International Conference on Pattern Recognition, 1992, vol. 2, pp. 1–4.
34) W. Zong, G. Huang, and Y. Chen, “Weighted extreme learning machine for imbalance learning,” Neurocomputing, vol. 101, pp. 229–242, 2013.
35) D. Vanderbult, “Soft self-consistent pseudopotentials in a generalized eigenvalue formalism,” Physical Review B, vol. 41, no. 11, pp. 7892, 1990.
36) H. Chen, A. Gallagher, and B. Girod, “Describing clothing by semantic attributes,” in Proceedings of European Conference on Computer Vision, 2012, pp. 609–623.
37) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
38) E. Gibaja and S. Ventura, “A tutorial on multilabel learning,” ACM Computing Surveys, vol. 47, no. 3, pp. 52, 2015.
39) S. Vittayakorn, K. Yamaguchi, A. Berg, and T. Berg, “Runway to realway: Visual analysis of fashion,” in Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 951–958.
40) S. Hidayati, C. You, W. Cheng, and K. Hua, “Learning and
recognition of clothing genres from full-body images," *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1647–1659, 2017.

41) S. C. Hidayati, K. Hua, Y. Tsao, H. Shuai, J. Liu, and W. Cheng, “Garment detectives: Discovering clothes and its genre in consumer photos,” in *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 471–474.

42) T. Durand, T. Mordan, N. Thome, and M. Cord, “Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 642–651.

43) M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?-weakly-supervised learning with convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

44) A. Mary and D. Dharma, “Coral reef image/video classification employing novel octa-angled pattern for triangular sub region and pulse coupled convolutional neural network (PCCCNN),” *Multimedia Tools and Applications*, vol. 77, no. 24, pp. 31545–31579, 2018.

45) A. Mary and D. Dejey, “Classification of coral reef submarine images and videos using a novel Z with tilted Z local binary pattern (ZTZLBP),” *Wireless Personal Communications*, vol. 98, no. 3, pp. 2427–2459, 2018.

46) A. Mary and D. Dharma, “Coral reef image classification employing improved LDP for feature extraction,” *Journal of Visual Communication and Image Representation*, vol. 49, pp. 225–242, 2017.

47) A. Nielsen, “Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data,” *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

48) X. Zhang, N. Guan, Z. Luo, and L. Lan, “Discriminative locality preserving canonical correlation analysis,” in *Proceedings of Chinese Conference on Pattern Recognition*, 2012, pp. 341–349.

49) J. Yang and X. Zhang, “Feature-level fusion of fingerprint and finger-vein for personal identification,” *Pattern Recognition Letters*, vol. 33, no. 5, pp. 623–628, 2012.

50) G. Lee, A. Singanamalli, H. Wang, M. Feldman, S. Master, N. Shih, E. Spangler, T. Rebeck, J. Tomsaszewski, and A. Madabhushi, “Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 284–297, 2015.

**Kazaha Horii** received his B.S. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2018. He is currently pursuing a Master degree at the Graduate School of Information Science and Technology, Hokkaido University. His research interests are multimodal processing and its applications. He is a student member of the IEEE.

**Keisuke Maeda** received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2015, 2017 and 2019, respectively. He is currently pursuing a Postdoctoral Fellow degree at the Faculty of Information Science and Technology, Hokkaido University. He is currently a Research Fellow of the Japan Society for the Promotion of Science. His research interests are multimodal processing and its applications. He is a student member of the IEEE and IEICE.

**Takahiro Ogawa** received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 2003, 2005, and 2007, respectively, all in electronics and information engineering. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the ITE Transactions on Media Technology and Applications. He is a member of the ACM, EURASIP, IEICE, and Institute of Image Information and Television Engineers (ITE) and a senior member of the IEEE.

**Miki Haseyama** received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 1986, 1988, and 1993, respectively, all in electronics. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEICE and Information Processing Society of Japan (IPSJ), a senior member of the IEEE and a fellow of Institute of Image Information and Television Engineers (ITE). She has been a Vice-President of the ITE, the Editor-in-Chief of the ITE Transactions on Media Technology and Applications, the Director of the International Coordination and Publicity of the IEICE.
Fig. 5 Confusion matrices of image classification results. A parameter of threshold value \((\text{Th})\) was set to a value which provided the best performance for each method in the test phase.
Fig. 6  Relationship between threshold value ($Th$) and final classification performance (F-measure). The red broken line represents the ideal of PM, and the blue broken line represents the ideal of CM1. Since there is no threshold parameter in CM2 and CM4, the F-measures are constant. PM ideal, CM1 ideal and CM2 correspond to “VF and canonical features projected GT-Att”, “VF and GT-Att” and “VF” in Table. 5, respectively.

Fig. 7  Results (Accuracy, Macro-P, Macro-R, Macro-F and Micro-F) of attribute estimation by each threshold value ($Th$). Since the method used for attribute estimation in PM, CM1 and CM3 are the same, the results of attribute estimation are the same. Note that the higher the value is, the better is the performance.
Fig. 8  Results (Hamming loss) of attribute estimation by each threshold value. Since the method used for attribute estimation in PM, CM1 and CM3 are the same, the results of attribute estimation are the same. Note that the lower the value is, the better is the performance.

Fig. 9  Some attributes estimated by PM with $Th = 0.3$. Results of estimation of representative attributes. White blocks are attributes estimated to be given and black blocks are attributes estimated not to be given.