LETTER *Special Section on Smart Multimedia & Communication Systems*

Multi-Task Convolutional Neural Network Leading to High Performance and Interpretability via Attribute Estimation

Keisuke MAEDA[†a)], ***Member***, **Kazaha HORII**[††], ***Nonmember***, **Takahiro OGAWA**[†††], ***and*** **Miki HASEYAMA**[†††], ***Members***

SUMMARY A multi-task convolutional neural network leading to high performance and interpretability via attribute estimation is presented in this letter. Our method can provide interpretation of the classification results of CNNs by outputting attributes that explain elements of objects as a judgement reason of CNNs in the middle layer. Furthermore, the proposed network uses the estimated attributes for the following prediction of classes. Consequently, construction of a novel multi-task CNN with improvements in both of the interpretability and classification performance is realized.

key words: *multi-task convolutional neural network, image classification, attribute estimation, interpretability*

1. Introduction

Convolutional neural networks (CNNs) are the most widely used deep learning methods in an image recognition field. CNNs have a large number of hidden layers, and lower convolution layers capture ordinary phrasing basic features (e.g., colors and shapes), and top layers near outputs of the network are able to learn more complicated structures [1]. Since the large number of hidden layers of CNNs transform visual features to discriminant features, CNNs realize highly accurate image classification [2]. However, there is no clear understanding of why CNNs perform so well [3]. For instance, since a CNN is an end-to-end training system, observable information is only input images and output classes. In other words, there is still little insight into the internal operation and behavior for human interpretaition of CNNs. This problem causes a distrust of classification results via CNNs. Therefore, it is necessary to provide an interpretation of CNNs.

Many interesting approaches to better understand classifier decisions and to gain insight into how CNNs operate have been proposed [4], [5]. Those approaches focused on region visualization that is related to classes. For instance, Selvaraju et al. [5] used the gradients of target classes flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the class. They could efficiently provide visual interpretability, and humans can identify the regions that contribute to classification results of CNNs. On the other hand, since humans often justify decisions verbally [3], it is expected that textual information that is more specific and understandable than visual information could give more obvious reasons to humans for the CNN-based classification results.

Generally, objects are identified on the basis of verbal descriptions called attributes that explain their elements [6]. Since attributes correspond to high-level properties of the objects that can be detected by machines and that can be understood by humans [7], attributes have attracted much attention in visual recognition research due to the fact that attributes provide detailed knowledge about the objects in images [1], [8]. Thus, attributes can improve not only object recognition performance but also interpretability of the classification results. Since attributes are verbal descriptions, they are important for understanding object appearance that provides interpretation of CNNs without ambiguity.

In order to realize attribute estimation and image classification simultaneously, we should consider a multi-task problem. Multi-task CNNs, which are trained by using multiple information such as classes and attributes related to the classes, are used in an image recognition field [1], [9]. General multi-task CNNs have multiple output layers as final results such as classes and attributes and they optimize multiple loss functions simultaneously [1], [9]. Since estimated attributes cannot directly affect the results of image classification, the image classification performance of multi-task CNNs tends to be reduced compared to that of single-task CNNs for predicting only classes. This is the drawback of general multi-task CNNs. In order to solve the above problem, we focus on the fact that attributes correspond to high properties of an object. In order to utilize attributes as semantic information for image classification, we estimate attributes in the middle layer of the network for image classification. It is expected that the use of estimated attributes enable interpretability to be improved while maintaining a level of performance close to that of single-task CNNs.

In this letter, we propose a novel interpretable multi-task CNN with attribute estimation for image classification. We add a layer for attribute estimation to a middle layer of a general CNN in order to provide interpretable results. Moreover, in order to enhance classification performance, the proposed method uses the estimated attributes as features

Manuscript received February 6, 2020.
Manuscript revised May 24, 2020.

[†]The author is with the Office of Institutional Research, Hokkaido University, Sapporo-shi, 060-0808 Japan.
[††]The author is with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
[†††]The authors are with the Faculty of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
a) E-mail: maeda@lmd.ist.hokudai.ac.jp

DOI: 10.1587/transfun.2020SML0006

in addition to the visual features in the CNN architecture. Although our method trains attributes and classes simultaneously like general multi-task CNNs, we focus on the relationship between attributes and classes and use attributes estimated by intermediate layers' output for image classification. This is the main contribution of our study. Our method realizes an interpretable multi-task CNN while maintaining a level of image classification performance close to that of single-task CNNs.

2. Interpretable CNN

Figure 1 shows an interpretable CNN architecture. Since CNNs gradually transform low-level features to high-level features by hidden layers and since attributes have potentials to bridge between low-level features and high-level class information [1], our method additionally constructs a full-connected layer for attribute estimation that is a multi-label problem on the intermediate layer of the CNN. Then, unlike a conventional multi-task CNN, our method performs image classification by full-connected layer's outputs based on both of the intermediate layer's outputs and the estimated attributes. Since attributes represent elements of the image and bridge between the visual features and the class information, the use of estimated attributes is effective for improving not only the interpretability of image classification results but also image classification performance.

Training images n $(= 1, \cdots, N$; N being the number of training samples) with classes $\boldsymbol{y}_n^{(c)} \in \{0,1\}^{D_y^{(c)}}$ and attributes $\boldsymbol{y}_n^{(a)} \in \{0,1\}^{D_y^{(a)}}$ ($D_y^{(c)}$ and $D_y^{(a)}$ being the numbers of classes and attributes, respectively) are given. Our method outputs a probability value $\boldsymbol{o}_n^{(a)} \in \mathbb{R}^{D_y^{(a)}}$ from the added full-connected layer for attribute estimation. Our method classifies images based on the intermediate layer's outputs and the estimated attributes. Our method obtains new features $\boldsymbol{z}_n = [\boldsymbol{x}_n^\top, \boldsymbol{o}_n^{(a)\top}]^\top \in \mathbb{R}^{D_x + D_y^{(a)}}$ by performing vector combination of the intermediate layer's output $\boldsymbol{x}_n \in \mathbb{R}^{D_x}$ and the probability value vector of the attribute estimation $\boldsymbol{o}_n^{(a)}$.

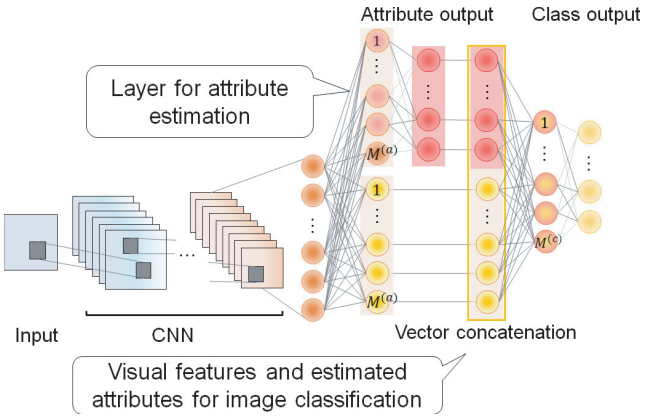

Fig. 1 Architecture of the proposed method. Our method adds a full-connected layer for attribute estimation to a CNN in order to provide interpretation of the CNNs. Then our method performs the final image classification by using the middle layer's outputs and the estimated attributes.

Our method outputs a probability value $\boldsymbol{o}_n^{(c)} \in \mathbb{R}^{D_y^{(c)}}$. In the training phase, the multi-task CNN is trained by using a loss function $L = \alpha L^{(c)} + (1-\alpha) L^{(a)}$, where $L^{(c)}$ is a categorical cross-entropy loss function for classes, and $L^{(a)}$ is a binary cross-entropy loss function for attributes. Furthermore, α is a trade-off parameter. Unlike a single-task CNN, introduction of the second term in the loss function is a technical point of our method.

In the test phase, given a test image, our method calculates an intermediate layer's output vector $\boldsymbol{x}$ and a probability value vector $\boldsymbol{o}^{(a)}$ for the attribute estimation. Then our method estimates attributes corresponding to the elements for which probability values exceed pre-determined threshold values ($\boldsymbol{t} = [t_1, \cdots, t_{D_y^{(a)}}]$). Then our method obtains a new feature $\boldsymbol{z}$ in the same manner as the training phase and performs image classification based on the class corresponding to the elements for which the probability value is maximum in a probability value vector $\boldsymbol{o}^{(c)}$ for the final image classification.

3. Experiment

In order to verify the effectiveness of our method (PM), we used Deep Fashion Database [10]. We used images for "Tank", "Blouse", "Shorts" and "Skirts", and these types of clothing were used as classes. Deep Fashion Database contains 50 categories, and the number of images in each category is much different, e.g., from tens to tens of thousands. Since the proposed method is a deep learning-based method and requires a large number of training images, we selected categories, which have more than 10,000 training images. In addition, to reduce class imbalance due to the bias in the number of training images between each category, we selected categories so that the number of images is balanced between selected categories. Thus, four classes were selected according to the above conditions. The number of images used in this experiment is shown in Table 1. Though this dataset has 1000 attributes, we selected K attributes based on mutual information between classes and attributes. We used the top K attributes with the highest degree of mutual information.

In our method, we used a pre-trained Inception-v3 [11], InceptionResNet-v2 [12] and Xception [13] models trained by ImageNet [2] as CNNs in order to verify the robustness of our method. We added a global average pooling layer and a full-connected layer with $M^{(a)}$ units to an intermediate layer of the CNNs for attribute estimation and added full-connected layers for image classification to the concatenated layer with $M^{(c)}$ units as shown in Fig. 1. We used the Adam optimization algorithm to train all parameters of the network with a batch size of 32. The learning rate was set to 0.001,

Table 1 Number of images used in our experiment.

	Blouse	Tank	Shorts	Skirts
Train	17752	11204	14195	10795
Validation	3389	2097	2700	2046
Test	3416	2128	2771	1933

and exponential decay rates for the moment estimates were set to 0.9 and 0.999, respectively. The trade-off parameter α and the number of attributes K were set to 0.5 and 200, respectively. The numbers of units of the added full-connected layers $M^{(a)}$ and $M^{(c)}$ were set to 1024 and 512, respectively. We evaluated the performance of the final image classification by using accuracy. For attribute estimation, we evaluated the performance by using micro accuracy. We used a general multi-task CNN according to [9] (CM) and Ideal in the experiment. Specifically, when constructing CM, it shares the layer just before the output layer of each CNN model in the same manner as PM. That is, the CM has two kinds of $M^{(a)}$ dimensional outputs for attribute estimation and image classification. Note that, although CM performs both attribute estimation and image classification, CM does not use estimated attributes for image classification. This is the difference between PM and CM. Since CM outputs estimated attributes for interpretation of the CNN and trains the network based on multi-task learning, CM is an appropriate method as a comparative method. Ideal performs only image classification, that is, general image classification. Ideal constructs a full-connected layer to pre-trained CNNs for the image classification. Since Ideal focuses only on image classification, the result of Ideal is the upper limit of the image classification performance. Thus, Ideal shows the upper limit of image classification by the CNNs. Note that both methods were calculated by using Geforce RTX2080Ti, and the required memory is less than 11GB.

Figure 2 shows results of the final image classification and the attribute estimation. Since PM is superior to CM in all network structures, the use of estimated attributes is effective for final image classification. The accuracy of PM is very close to that of Ideal for Inception-v3 and InceptionResNet-v2. Moreover, the micro accuracy of the attribute estimation is very high. For InceptionResNet-v2 and Xception, the micro accuracy of PM is superior to that of CM. Therefore, our method can provide interpretability of the CNN and realizes more accurate image classification than do general multi-task CNNs while maintaining a level of image classification performance close to that of single-task CNNs. Furthermore, we show numerical accuracy of image classification in PM and CM. Accuracy values of PM for Inception-v3, InceptionResNet-v2 and Xception are 0.883, 0.874, 0.811, respectively. Also, those of CM are 0.872, 0.845, 0.796, respectively. Furthermore, in order to verify the statistically significant difference, we applied Welch's t-test to the results obtained by Inception-v3 of five trials since the accuracy of Inception-v3 has the slightest difference between PM and CM. Thus, it was confirmed that we have a sufficient statistical advantage (p-value= 0.022).

Figure 3 shows examples of correctly classified images with estimated attributes by PM. PM not only correctly classifies images but also provides an interpretation by estimating attributes that correspond to high properties of the object. Therefore, by outputting the estimated attributes in the intermediate layer of the CNN, we can understand how the CNN estimated. Note that since Ideal cannot provide interpretation, attribute estimation is more effective for providing interpretablility into a CNN.

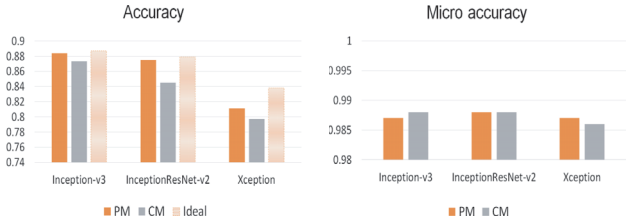

Fig. 2 Results of final image classification (accuracy) and attribute estimation (micro accuracy). Note that Ideal that focuses only on image classification is the upper limit of the image classification performance.

(a) Class: Skirt. Attributes: mini, sequined

(b) Class: Blouse. Attributes: chiffon

(c) Class: Short. Attributes: crochet, lace, crochet lace

(d) Class: Tank. Attributes: crop

Fig. 3 Examples of correctly classified images with estimated attributes.

Furthermore, since the number of layers and neurons are almost the same as those of PM, the training time is almost the same. Specifically, the training time of PM for Inception-v3, InceptionResNet-v2 and Xception is 3,628 (sec), 7,208 (sec) and 4,950 (sec), respectively, and that of CM is 3,647 (sec), 7,191 (sec) and 4,925 (sec), respectively. Thus, since PM leads to high performance and interpretability although there is no different between PM and CM in terms of the computation cost, the effectiveness of PM is verified.

4. Conclusion

We have proposed a multi-task convolutional neural network leading to high performance and interpretability via attribute estimation. The proposed method performs attribute estimation and image classification simultaneously based on our novel multi-task CNNs. Specifically, our method uses the estimated attributes for image classification in order to maintain a level of image classification performance close to that of single-task CNNs. Moreover, the estimated attributes provide interpretation of the classification results. The effectiveness of our method was verified by an experiment. On the other hand, since this method can be applied to data with images and attributes, it is necessary to perform accuracy verification using various data other than clothing data as future work in order to confirm the generalization performance of our method.

Acknowledgments

This work was partly supported by the MIC/SCOPE #181601001.

References

[1] A. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," IEEE Trans. Multimedia, vol.17, no.11, pp.1949–1959, 2015.
[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] Z. Lipton, "The mythos of model interpretability," arXiv preprint arXiv:1606.03490, 2016.
[4] P. Kindermans, K. Schütt, M. Alber, K. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: PatternNet and PatternAttribution," arXiv preprint arXiv:1705.05598, 2017.
[5] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Proc. IEEE Int. Conf. Computer Vision, pp.618–626, 2017.
[6] C. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.3, pp.453–465, 2014.
[7] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.819–826, 2013.
[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," Proc. the IEEE Conf. Computer Vision and Pattern Recognition, pp.1778–1785, 2009.
[9] Y. Takada, N. Inoue, T. Yamasaki, and K. Aizawa, "Similar floor plan retrieval featuring multi-task learning of layout type classification and room presence prediction," Proc. IEEE Int. Conf. Consumer Electronics, pp.1–6, 2018.
[10] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016.
[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015.
[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," Proc. AAAI Conf. Artificial Intelligence, pp.4278–4284, 2017.
[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1251–1258, 2017.