Generating Visual and Semantic Explanations with Multi-task Network

Wenjia Xu[1,2](✉), Jiuniu Wang[1,2], Yang Wang[1], Yirong Wu[1,2], and Zeynep Akata[3]

[1] Department of Electrical Engineering, University of Chinese Academy of Sciences, Beijing, China
xuwenjia16@mails.ucas.ac.cn

[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China

[3] Cluster of Excellence Machine Learning, University of Tübingen, Tübingen, Germany

Abstract. Explaining deep models is desirable especially for improving the user trust and experience. Much progress has been done recently towards visually and semantically explaining deep models. However, establishing the most effective explanation is often human-dependent, which suffers from the bias of the annotators. To address this issue, we propose a multitask learning network (MTL-Net) that generates saliency-based visual explanation as well as attribute-based semantic explanation. Via an integrated evaluation mechanism, our model quantitatively evaluates the quality of the generated explanations. First, we introduce attributes to the image classification process and rank the attribute contribution with gradient weighted mapping, then generate semantic explanations with those attributes. Second, we propose a fusion classification mechanism (FCM) to evaluate three recent saliency-based visual explanation methods by their influence on the classification. Third, we conduct user studies, quantitative and qualitative evaluations. According to our results on three benchmark datasets with varying size and granularity, our attribute-based semantic explanations are not only helpful to the user but they also improve the classification accuracy of the model, and our ranking framework detects the best performing visual explanation method in agreement with the users.

Keywords: Multi-task learning · Explainable AI

1 Introduction

Deep learning has led to remarkable progress in computer vision tasks. However, despite their superior performance, the black-box nature of deep neural networks harms user trust. In order to build interpretable models that can explain their behaviour, previous works visually point to the evidence that influences the network decision [17,24], or provide semantic explanations that justify a

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12535, pp. 620–635, 2020.
https://doi.org/10.1007/978-3-030-66415-2_40

category prediction [11,12]. When it comes to generating visual explanations, example methods includes visualizations via gradient flow or filter deconvolution [29,30,39], visualizing class activation maps [28,42] and measuring the effect of perturbations on input images [7,26,27].

However, there is no unified evaluation metric to determine the most effective visualization technique. Although user-studies are widely used to judge the effectiveness of visualization methods, it is unscalable since humans are not on-demand software that can be employed at anytime. Among automatic evaluation methods, RISE [26] proposes to evaluate the influence of different regions to the decision maker by deleting/inserting pixels. However, this requires insertion/removal of many pixel combinations for each image, e.g., several iterations are needed for evaluating one image, which is time consuming.

Natural language explanations [11] is a complementary way to justify neural network decisions. These explanations are usually generated by feeding images and predicted class labels into LSTM [10]. A drawback is that the semantic explanations may lack the class discriminative ability, missing essential details to infer the image label [20].

In this work, our primary aim is to generate visual and semantic explanations that are faithful to the model, then quantitatively and objectively evaluate the justifications of a deep learning based decision maker. To realize this aim, we propose a visual and semantic explanation framework with an integrated quantitative and objective evaluation mechanism without requiring the user in its training or inference steps. We classify and embed attributes to help the category prediction. The semantic explanation is generated based on gradient weighted mapping of the attribute embedding, then evaluated on its image and class relevance. Furthermore, to evaluate the visual explanation methods in this framework, we propose a fusion classification mechanism. The input image is filtered by its visual explanation map and then fed into a classifier. We evaluate the methods based on the classification accuracy.

We argue that an explanation is faithful to the model it is interpreting, if it can help to improve the performance of that black box model. For instance, for the task depicted in Fig. 1, an accurate visual explanation model should attend to the clothing related regions to recognize a "Clothing Store". Hence, if the background and non-relevant pixels are weakened and the clothing related regions are preserved, e.g., the image is filtered through the attention mechanism, the classification result should not be degraded. The same holds for attribute-based justifications. For instance, an accurate semantic explanation of the predicted label "clothing store" should inform the user about the most discriminative attributes such as "enclosed area, cloth, indoor lighting". The effectiveness of this explanation can be verified by feeding these attributes to the network for the classification task.

Our main contributions are summarized as follows. (1) We propose a visual and semantic explanation framework that generates explanations that are fidelity to the model. (2) We design an integrated evaluation mechanism with quantitative and qualitative evaluations as well as user study on the explanations. The quantitative evaluation is automatically performed according to the influence of explanations on classification tasks. (3) We showcase on three datasets that our semantic explanations are faithful to the network they are interpreting. Three representative visual explanations, i.e. Grad-Cam [28], Back-Propagation [29] and RISE [26] are evaluated and the quantitative results agree with the user preference.

2 Related Work

In this section, we summarize the prior work on multitask learning and explainability research related to ours.

Multitask Learning. Multitask learning is a popular method that enables us to train one neural network to do many tasks. Some prior works have shown that learning multiple tasks can improve the generalization of the network and give better performance than doing these tasks separately [2]. For instance, a multitask network for segmentation can improve the performance of object detection while being much faster [4]. In our work, we train a network on image recognition and attributes classification simultaneously, motivated by the fact that sharing lower-level features can benefit these two tasks and result in better performance.

Textual Explanation. Generating semantic explanations has gained interest in the community. Among those, Hendricks et al. [11,12] take the image and its predicted category label as input, and generate explanations with a conditioned LSTM [10]. Although these explanations build a sound basis for enabling user acceptance of the deep models and improving user trust, they cannot guarantee the fidelity to the model. The conditioned LSTM model trained on human annotated captions may generate sentences describing the image content, rather than the real reason for the network decision.

We take advantage of semantic attributes to generate textual explanations. Attributes are human-annotated discriminative visual properties of objects [18,25]. In zero-shot learning [18,36,37], they are used to build intermediate representations to leverage the lack of labeled images as the model does not have access to any training examples of some classes. [16] and [15] apply attributes as a linguistic explanation for video and image classification. They select the attributes by its interaction information with the input images. Our method differs in that we define the important attributes by how much the they influence the classification task, and aggregate them into a semantic explanation. Image attributes are used to boost the performance for fine-grained classification [6,41], face recognition [13], and image-to-text generation [41]. If annotated on a per-class basis, attributes are both effective and cheap [33].

Visual Explanation. We distinguish between two types of visual explanations: interpretation models and justifying post-hoc reasons. The former visualize the filters and feature maps in CNNs, trying to interpret the knowledge distilled by the model [22,38,39]. For instance, [39] applies a Deconvolutional Network (DeConvNet) to project the feature activations back to the input pixel space. The latter determines which region of the input is responsible for the decision by attaching importance to pixels or image regions. Gradient-based methods back-propagate the loss to the input layer [29,30]. Although these methods generate high-resolution details, they can not localize the image area that the target category focuses on. Visualizing linear combination of the network activations and incorporating them with class-specific weights is another direction [28,42]. Class Activation Mapping (CAM) [42] and its extension Grad-CAM [28] produce class-specific attention maps by performing a weighted combination on forward activation maps. On the other hand, model-agnostic methods propose to explain models by treating them as black boxes. Perturbation-based methods manipulate the input and observe the changes in output [7,8,26,27]. A linear decision model (LIME) [27] feeds super-pixel-masks into the black box and generates attention maps. An extension of LIME, RISE perturbs the input image with random masks and generates weights with the output probabilities, then produce the attention map by the weighted combination of random binary masks. [8] and [7] extends the perturbation to a trainable parameter and generates more smooth masks.

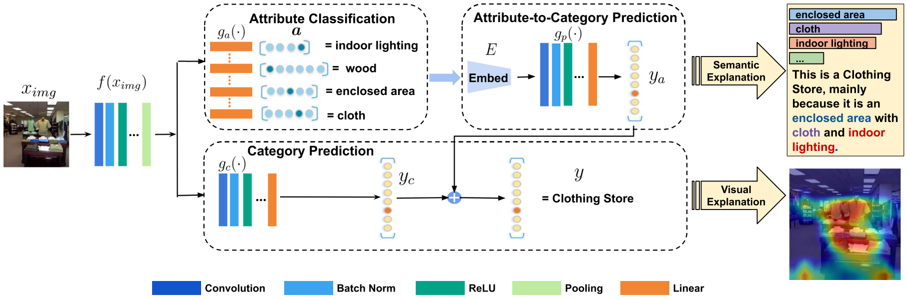


Fig. 1. Overview of our multitask learning network (MTL-Net). $f(x_{img})$ is the feature extraction network in our model. The network contains two pipelines. Category prediction network predicts the label y_c for an input image x_{img}. Attribute classification network predicts the attributes a. Attribute-to-Category prediction network infers the category y_a by attributes embedding. Then the attributes with high contribution to y_a are aggregated into a template-based explanation. Saliency-based visual explanation reflects the attention of y_c on the input image.

Evaluating Explanations. Although visual explanations are an intuitive way of understanding the internal thought process of a neural network, it is not trivial to measure the effectiveness of the visualization method. In recent years, various evaluation methods are performed. The most widely accepted measure of effectiveness is user studies [28,42]. Some explanation methods perform quantitative evaluation methods such as Pointing ame [19,28,40], sanity checks [1], and Deletion-insertion [26].

Since visual explanations reflect the salient area that activates the feature map, improving the visualization would be beneficial for classification. Hence, we propose to fuse the input image with the explanation maps and then measure how the classification accuracy is influenced. The testing procedure is processed only once when evaluating the explanation methods.

3 Visual-Semantically Interpretable Framework

In this section, we introduce how we integrate the attribute prediction with image classification. Then we detail how to generate semantic explanations via attribute contribution. Finally, we present visual explanations generated by various visualization methods and evaluate them with the fusion classification mechanism.

3.1 Multitask Learning Network

Learning multiple complementary tasks would improve the generalization capability of the network, and improve the accuracy of predictions compared to performing these tasks separately. Deep neural networks for image classification uses category level labels as the supervision signal [32,35]. While attributes reveal essential characteristics of objects complementary to image classes [37]. In our multitask learning network (MTL-Net), we combine three modules regarding category and attribute classification within a unified framework as shown in Fig. 1. They are category prediction, attribute classification and attribute-to-category prediction. Given an input image x, our task is to predict the image label y as well as the attribute a with the following steps.

In category prediction, given input image x_{img}, we first extract the image feature v, then we pass these image features into a linear classifier and get the predicted result y_c:

$$v = f(x_{img})\,, y_c = g_c(v)\,. \tag{1}$$

To predict N_a attributes of one image, we apply a linear classifiers to learn the i_{th} attribute $a_i \in \mathbb{R}^{d_{a_i}}$ from the image feature v:

$$a_i = g_a^i(v)\,. \tag{2}$$

While predicting attributes will help the image encoder $f(\cdot)$ to extract semantic information regarding the attributes and help the image classification process, the predicted attributes are not contributing to the image classification directly. Thus we combine them in the attribute-to-category prediction. We first follow the word embedding method [9] to embed a into a matrix $E \in \mathbb{R}^{N_a \times d}$:

$$E = \mathit{concat}[a_1 \cdot W_1, a_2 \cdot W_2 \cdots, a_{N_a} \cdot W_{N_a}]\,, \tag{3}$$

where W_i is the embedding matrix with dimension $\mathbb{R}^{d_{a_i} \times d}$, and N_a is the number of attributes. Then the embedding E is feed into a linear classifier to get the predicted result y_a:

$$y_a = g_p(E)\,. \tag{4}$$

Thus, the final class prediction is

$$y = y_c + \alpha \cdot y_a\,, \tag{5}$$

where α is a hyper parameter.

We optimize the MTL-Net with the cross-entropy loss L between predicted class and the ground truth y_{gt}:

$$L = CE(y, y_{gt})\,. \tag{6}$$

In order to align the predicted attributes with interpretable semantic meaning, we propose to optimize the attribute classification module with human annotated class attributes $\mathcal{A} \in \mathbb{R}^{N_a}$. The attribute classification network would be optimized according to the objective:

$$L_{attri} = \frac{1}{N_a}\sum_{i=1}^{N_a} CE(\mathcal{A}_i, a_i)\,. \tag{7}$$

Our final loss L is the weighted combination of the above three loss: L and L_{attri}:

$$L = L + \beta \cdot L_{attri}\,. \tag{8}$$

3.2 Interpreting MTL-Net

Here we detail our method for selecting the attributes that make an important impact on the results, and evaluating the image area that the network pays most attention to. We propose a gradient weighted mapping to figure out the attribute contribution in image classification, and generate language explanations using a predefined template. Furthermore, we apply various visualization methods, i.e. Back-propagation (BP) [29], Grad-CAM [28] and RISE [26], to generate saliency maps that provide visual justifications of the network classification process. Then we evaluate them with our fusion classification mechanism.

Generating Attribute-Based Explanations. In attribute-to-category prediction, the predicted image category y_a is inferred by the prediction of the N_a attributes for every image. In order to determine which attributes are most important to the score y_a^k for the predicted class k, we use gradient mapping to generate a saliency map $M^a \in \mathbb{R}^{N_a \times d}$ for attributes embedding E. As is discussed in [30], we only consider the positive impact on increasing the predicted class score, thus the attribute contribution of the i-th attribute a_i is determined as the sum of positive values in $M_i^a \in \mathbb{R}^d$:

$$C_{a_i} = \sum_{j=1}^{d} \mathbb{1}_{M_{ij}^a > 0} \cdot M_{ij}^a, \qquad \text{where } M_{ij}^a = \frac{\partial y_a^k}{\partial E_{ij}} \,. \tag{9}$$

Here M_i^a denotes the gradient of the score y_a^k on the attribute embedding.

The attribute contribution C_{a_i} $(i = 1, \ldots, N_a)$ indicates which attributes have more positive impact on predicting class k. Consequently, we rank the score and pick the attributes with the highest contribution C_{a_i}. And we select the top three attributes to form our semantic explanation.

Generating Visual Explanations. To generate visual explanations and evaluate which one is more fidelity to the network, we utilize three representative methods:

BP. Back-Propagation (BP) [29] computes the saliency map by back propagating the output score y_c into the input layer. Based on the gradient for each pixel (i, j) of the input image, the saliency map is computed as:

$$M_{ij} = max_l \left| \frac{\partial y_c}{\partial x(i,j,l)} \right| \,, \tag{10}$$

where $max_l \left|\cdot\right|$ denotes the maximum magnitude across all color channels l.

Grad-CAM. Grad-CAM [28] uses the gradient flowing back to a specific convolutional layer to calculate the importance weight α_k for every feature map A^k in that layer,

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \,. \tag{11}$$

And the final saliency map M is a weighted combination of the feature maps,

$$M = \mathrm{ReLU}(\sum_k \alpha_k A^k) \,, \tag{12}$$

where ReLU means that Grad-CAM only focuses on the features that have a positive influence on network output. And M is resized to the size of input image when we use it.

RISE. For each input image x, RISE [26] generates numerous masks $M^{(i)}$ to cover x. The author assumes that the output score of the whole network for masked image $F(x \odot M^{(i)})$ reflects the importance of that mask, where $\odot$ denotes element-wise multiplication. Thus the final attention map is the weighted sum of these masks,

$$M = \frac{1}{\mathbb{E}[M] \cdot N} \sum_{i=1}^{N} F(x \odot M^{(i)}) \cdot M^{(i)} \,, \tag{13}$$

where $\mathbb{E}[M]$ denotes the expectation of masks, and N is the number of generated masks.

Evaluating Visual Explanations. The lack of objective evaluation metrics for the performance of these visualization methods may hinder user acceptance. We conjecture that visual justifications would be trustworthy for the user if they improve the performance of the black box neural network on the task that they are visualizing. Hence, we propose image classification as a task to objectively evaluate the visual justification methods without human annotation.

Saliency maps indicate the importance of each pixel and retain the same spatial information as input images. We propose a fusion classification mechanism (FCM), where we overlay the saliency map M onto the raw image x_{img}, and generate the filtered image x_{fuse}. So that visual justifications can be evaluated automatically by training and testing our network on those fused images. We normalized the explanation maps into $[0, 1]$, to make equal compare among every explanation maps. The overlay method is described as,

$$x_{fuse} = (M + \lambda) \odot x_{img} \,, \tag{14}$$

where λ is a constant parameter that determines how much image content is least preserved, and when $\lambda = 0$ there might be image pixels being removed directly. $\odot$ denotes element-wise multiplication. We then feed the fused image x_{fuse} into the multitask learning network as shown in Fig. 1. Finally, we rank the saliency models based on their performance in classification. The ranking shows us that visual explanation models lead to a higher accuracy can capture important image regions for predicting the right answer.

4 Experiments

In this section, we start by introducing the dataset. We then present our results that validate the proposed multitask learning network (MTL-Net) on three datasets, indicating consistent improvements on single-task networks. Furthermore, our predicted attributes and their aggregated semantic explanations are presented and evaluated. The visual explanations are generated by three well-known visualization methods, and our proposed evaluation technique validates their effectiveness and ranks them based on the class prediction performance.

Datasets. We use three datasets for experimental analysis. CUB-200-2011 (CUB) [34] is a fine-grained dataset for bird classification, with 11,788 images from 200 different types of birds. The dataset consists of 312 binary attributes

Table 1. Ablation study for different settings in MTL-Net. We report the results for baseline models SE-ResNeXt-50 [14], Inception-v4 [31] and PNASNet-5 [21] on y_c. CP represents the accuracy of category prediction y_c trained together with attribute classification, and A2CP represents the accuracy of y when combining category prediction and attribute-to-category prediction. We also report the accuracy for attribute classification in MTL-Net.

Models	SUN	CUB	AwA
SE-ResNeXt-50 [14]	38.28	74.30	94.74
PNASNet-5 [21]	42.53	83.20	95.47
Inception-v4 [31]	35.49	78.90	94.22
CP (ours)	44.70	83.44	**95.71**
A2CP (ours)	**44.90**	**83.77**	95.61
Attribute classification	93.07	88.12	99.03

that describe the color, shape and other characters for 15 body part locations. SUN Attribute Database (SUN) [25] is a fine-grained scene categorization dataset, and consists of 14,340 images from 717 classes (20 images per class). Each image is annotated with 102 attributes that describe the scenes' material and surface properties. Animals with Attributes (AwA) [18,36] is a coarse-grained dataset for animal classification, containing 37,322 images of 50 animal classes. 85 per-class attribute labels [23] are provided in the dataset, describing the appearance and the living habits of the animals.

Implementation Details. The baseline model in MTL-Net is PNASNet-5 [21] pretrained on ImageNet [5] and then finetuned on three datasets separately. The classifier g_i and g_a have the same structure: 2-layer CNN and one linear layer. We train our model with SGD optimizer [3] by setting $momentum = 0.9$, $weight\ decay = 10^{-5}$. We set $\alpha = 1$ and tuned β from 0.1 to 1.5 for different datasets. While evaluating saliency maps, we set λ as a matrix with each element equals to 0.3.

4.1 Evaluating Semantic Explanations

In this section, we quantitatively evaluate our attribute-based semantic explanations in two aspects: the fidelity to the model and the alignment with human annotation. Then we perform human study and qualitative analysis to discuss how well is the semantic explanations when making the network interpretable to users.

Quantitative Analysis. Here, we validate our multitask learning network (MTL-Net) on three benchmark datasets, i.e. CUB, AWA and SUN, for the image classification task.

We report the classification accuracy of category prediction y_c and the final result y, denoted as CP and A2CP respectively, in Table 1. As comparison, we choose the classification accuracy of SE-ResNeXt-50 [14], Inception-v4 [31] and PNASNet-5 [21] as baseline.

Introducing the attribute classification loss to the original image classification task improves the accuracy. As is shown in Table 1, we improve the accuracy of the baseline model on all datasets, achieving 44.70% on SUN, 83.44% on CUB and 95.71% on AWA. These results demonstrate that introducing attributes to image classification not only makes the models more explainable, e.g., predicting attributes such as "white crown, pink legs, white belly" is more informative than only predicting the category "slaty backed gull". After integrating the attribute-to-category prediction and category prediction, the classification accuracy is further improved.

Table 2. The user study for semantic explanations. Image relevance refers to the question: "Does the sentence match the image content?". Class Relevance refers to "Is the explanation reasonable for the prediction?". According to the user, our semantic explanations are image relevant, and reasonable.

Question	Options	Percentage
Image relevance	High	68.4%
	Somewhat	27.4%
	No	4.2%
Class relevance	Yes	68.6%
	No	31.4%

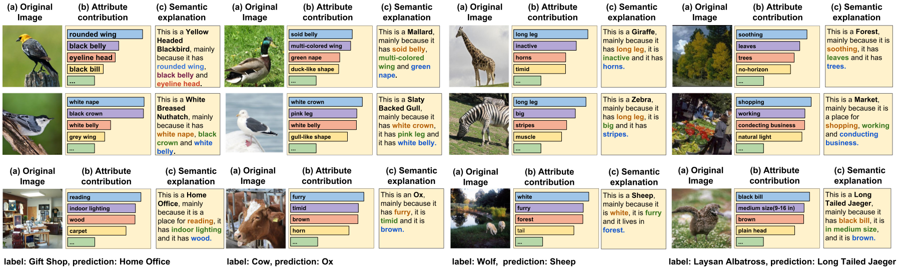
Table 3. The classification accuracy on three datasets (left) and the user study (right) for the fusion classification mechanism (FCM). No-VIS denotes the classification accuracy generated by our MTL-Net. Grad-CAM, BP, RISE refer to the FCM equipped with three visualization methods.

FCM	CUB	AwA	SUN	User study
No-VIS	83.77	95.61	44.90	N/A
Grad-CAM [28]	84.24	96.13	45.27	35.2%
BP [29]	81.87	93.86	41.73	26.2%
RISE [26]	**85.17**	**96.84**	**46.50**	**38.6%**

We also evaluate the predicted attributes by how well they are aligned with their semantic meaning. The predicted attributes in MTL-Net are compared with the class attributes annotated by a human, and we report the attribute classification accuracy on three datasets in Table 1. On average, the predicted attributes are agree with the human annotation, achieving an accuracy of 93.7%, 88.12% and 99.03 for three datasets. Note that the attributes in SUN and AwA dataset are binary, indicating the existence of the attribute. While the attributes in CUB dataset are multi-dimension, for instance, the head color attribute has fifteen options. That can explain why the accuracy of CUB dataset is slightly lower than the other two datasets.

User Study. Semantic explanations are mainly targeted towards the end user and aim to improve the user trust in the machine learning system. To determine if users find our semantic explanations trustworthy, we perform a user study on visual and semantic explanations. CUB being a fine-grained dataset, only bird experts can tell their difference. Hence, we selected 100 images from the scene categorization dataset SUN and the animal classification dataset AwA. Our user group is composed of five university graduates with an average age of 25. In this section, we present our results on semantic explanations for clarity, however, our user study on visual explanations presented in the following section is identical in the number, the demographics, the age and gender of the users as well as the number of images to be evaluated.

In the user study for semantic explanations, our aim is to evaluate two factors: if the semantic explanation is image relevant and how well can they help the user in understanding the black-box model. The annotators are given an image as well as a semantic explanation and are asked to answer two questions, i.e. "Does the sentence match the image?" and "Is the explanation reasonable for the prediction?". We present the results of this study in Table 2. For the question related to the image relevance, 68.4% of the attribute-based semantic explanations are marked as "highly related to the image", while only 4.2% of these results were found irrelevant by the users. For the question relevant to the credibility for the prediction, 68.6% of the sentences are found reasonable for the predicted label. These results complement our prior results and show that our attribute-based semantic explanations are image relevant and reasonable for the label that the model predicts.



This is a **Yellow Headed Blackbird**, mainly because it has rounded wing, black belly and eyeline head.

This is a **Mallard**, mainly because it has soid belly, multi-colored wing and green nape.

This is a **Giraffe**, mainly because it has long leg, it is inactive and it has horns.

This is a **Forest**, mainly because it is soothing, it has leaves and it has trees.

This is a **White Breased Nuthatch**, mainly because it has white nape, black crown and white belly.

This is a **Slaty Backed Gull**, mainly because it has white crown, it has pink leg and it has white belly.

This is a **Zebra**, mainly because it has long leg, it is big and it has stripes.

This is a **Market**, mainly because it is a place for shopping, working and conducting business.

This is a **Home Office**, mainly because it is a place for reading, it has indoor lighting and it has wood.

This is an **Ox**, mainly because it has furry, it is timid and it is brown.

This is a **Sheep**, mainly because it is white, it is furry and it lives in forest.

This is a **Long Tailed Jaeger**, mainly because it has black bill, it is in medium size, and it is brown.

label: Gift Shop, prediction: Home Office

label: Cow, prediction: Ox

label: Wolf, prediction: Sheep

label: Laysan Albatross, prediction: Long Tailed Jaeger

Fig. 2. Our MTL-Net predicts image category and attributes. We select the attributes having the highest contribution on the prediction label (b). The attributes are then aggregated into a template-based semantic explanation (c). The top two rows show our the explanations for right predictions, while the bottom row shows semantic explanations with wrong predictions. The "label" and "prediction" under each image indicates the ground truth label as well as the predicted label.

Qualitative Evaluation. In this section, we evaluate our attribute-based semantic explanations qualitatively, by looking at the sentences generated from the three highest ranked attributes together with the predicted label. Figure 2 shows two rows of example images with their predicted attributes where the label was correctly predicted. In the last row, we present three examples with their highest ranked attributes despite their wrong class predictions.

We observe from both the positive and the negative examples that our explanations correctly reflect the content of the image and the characters of the objects. For instance, in fine-grained bird classification results, our model correctly associates the attributes "green nape, multicolored wing, solid belly" with *Mallard* and "white crown, pink leg and white belly" with *Slaty Backed Gull*.

By looking at the attributes and the predicted label, a user can understand why this prediction was associated with these attributes. For instance, the explanation for zebra points out the most prominent attributes such as "long leg" and "stripe". While for a *Forest* image our model predicts the attributes "soothing, leaves, trees", and for a "Market" it associates "shopping, working and conducting business".

On the other hand, the users might find the reason for a wrong prediction by investigating the semantic explanations. For instance, we observe that due to "reading, indoor lighting, wood", an image for *gift shop* is wrongly predicted as *Home Office*. Arguably, the image looks more like a *home office* than a *gift shop*, i.e. correct class. Similarly, for the *wolf*, due to the unusual color of the animals (i.e., "white") and the tranquillity of the environment (i.e., "forest"), the label is predicted as *sheep*.

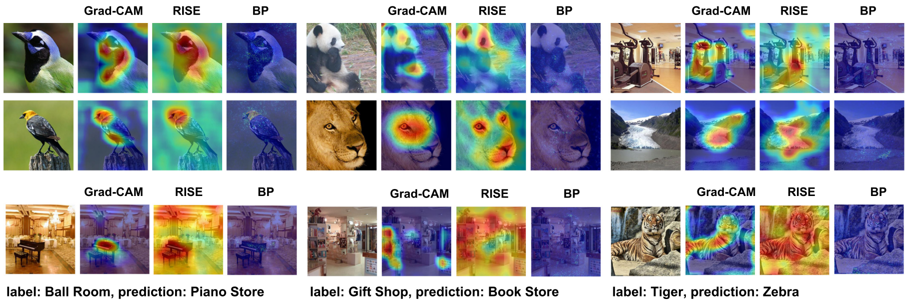

Grad-CAM
RISE
BP
Grad-CAM
RISE
BP
Grad-CAM
RISE
BP
Grad-CAM
RISE
BP
Grad-CAM
RISE
BP
Grad-CAM
RISE
BP
label: Ball Room, prediction: Piano Store
label: Gift Shop, prediction: Book Store
label: Tiger, prediction: Zebra

Fig. 3. Visual explanations of the correct labels generated by three methods, Grad-CAM, BP and RISE. Images on the left are from CUB, the middle from AWA and the right are from SUN datasets. The bottom row shows the visual explanations for wrong category prediction.

4.2 Evaluating Visual Explanations

To visually justify the classification decision of the model, we use three well-known visual explanation methods, Grad-CAM [28], BP [29], and RISE [26]. We compare them quantitatively in terms of the performance in fusion classification mechanism, through user studies, and qualitatively by visual inspection.

Visual Explanations in Image Classification. We evaluate the fidelity of the visual explanations generated by three models, by using them in the task they are interpreting, i.e. image classification. We first generate the attention maps of the predicted class in our MTL-Net, then fuse the image with the attention map on our fusion classification mechanism (FCM). Then we train and classify the fused image with MTL-Net again, and rank the visual explanations concerning the classification accuracy.

As presented in Table 3, our results indicate that images fused by saliency maps lead to slight improvements in the classification accuracy compared to the case with no fusion. Among the saliency-based explanation methods, RISE [26] consistently achieves better performance than BP and Grad-CAM. BP performs poorly on the fine-grained CUB dataset, and it may because the pixels that BP marks out are spread out all over the bird and they are not distinguishing between similar species (see Fig. 3). On the other hand, we experiment with all these models and indicate that Grad-CAM and BP results are much faster than RISE, since RISE requires multiple times of testing for every image.

User Study. Given three visual explanations, in this section, we aim to determine the visualization that is the most trustworthy for the users. We indicate that "All robots predicted the image as airfield", show visual explanations generated by Grad-CAM, RISE and BP to the annotators, and ask them to answer the question "Which robot do you trust more?". In this way, the annotators first evaluate if the label is correct by looking at certain regions in the image, and then they compare their attention with the visualizations, finally pick the one that matches their mental model. Using 100 images sampled from AWA and SUN datasets, we rank the saliency-based explanation models.

Our results shown in Table 3 (rightmost column) indicate that 38.6% of the annotators trust RISE, 35.2% of them trust Grad-CAM, and 26.2% of them vote for BP. This result is consistent with our fusion classification mechanism (FCM) for evaluating the quality of the saliency-based visualization. As a conclusion, if explanations are helpful for the users, they are expected to perform well in FCM. Although human study is a worthwhile and important evaluation criterion, it can be replaced by our automatic evaluation if time and labor limited.

Qualitative Evaluation. In this section, we evaluate the interpretability of the visual explanations in our MTL-Net. The first two rows in Fig. 3 show the visualizations for the correctly predicted images. In the last row, we present the visualization for images with the wrong prediction.

From both the results with correct and incorrect class prediction, we observe that Grad-CAM and RISE highlight important image regions to explain the network decision, while BP emphasizes a distributed set of pixels that are influential for the classification result. Hence, by looking at the masked image generated by Grad-CAM and RISE, one can easily figure out which part of the image the network focuses on for a particular decision. Generally, Grad-CAM offers a more concentrated focus due to the up-sample operation it takes. While RISE and BP consider more pixels when evaluating the importance.

With the negative predictions presented in the last row, the visual explanations can help the user to understand the causes for wrong predictions.

Indeed, most of the wrong predicted images may be confusing even for a human. For instance, when explaining why the image with the label *Ball Room* is predicted as a *Piano Store*, the visual explanation focuses on the piano and the indoor lighting. The wrong prediction of *Tiger* is an interesting example in that the attention map of Grad-CAM mainly focuses on the stripes that zebra also have. These results indicate that visual explanations can reveal show the weakness of network to the users, e.g., the network typically makes a mistake when it only focuses on wrong details, instead of considering the image as a whole. Moreover, RISE generates more scattered distribution for the wrong predictions. And that might be another clue for identifying wrong classified images.

5 Conclusion

In this work, we propose a visually and semantically interpretable multitask learning network. We introduce attributes into image category prediction and propose a new method to generate attribute-based semantic explanations intuitive for the user. Qualitative evaluations and the user study reveals that our semantic explanations are both class discriminative and image relevant. Moreover, we propose a quantitative evaluation technique to evaluate the effectiveness of visual explanations based on their performance in image classification. Future work includes investigating the network flaws with these explanations and further improve the network.

Acknowledgments. This work has been partially funded by the ERC grant 853489 - DEXIM (Z.A.) and by DFG under Germany's Excellence Strategy EXC number 2064/1 Project number 390727645.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, pp. 9505–9515 (2018)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: NIPS (2007)
3. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT 2010, pp. 177–186. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-7908-2604-3_16
4. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: IEEE CVPR (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE CVPR (2009)
6. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: IEEE CVPR (2012)
7. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV (2019)
8. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3429–3437 (2017)

9. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: NIPS (2016)
10. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005)
11. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 3–19. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_1
12. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 269–286. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_17
13. Hu, G., et al.: Attribute-enhanced face recognition with neural tensor fusion networks. In: IEEE ICCV (2017)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR (2018)
15. Kanehira, A., Harada, T.: Learning to explain with complemental examples. In: IEEE CVPR (2019)
16. Kanehira, A., Takemoto, K., Inayoshi, S., Harada, T.: Multimodal explanations by predicting counterfactuality in videos. In: IEEE CVPR (2019)
17. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self driving vehicles. In: ECCV (2018)
18. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE CVPR (2009)
19. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE TPAMI **36**, 453–465 (2014)
20. Li, Q., Fu, J., Yu, D., Mei, T., Luo, J.: Tell-and-Answer: towards explainable visual question answering using attributes and captions. In: EMNLP (2018)
21. Liu, C., et al.: Progressive neural architecture search. In: ECCV (2018)
22. Olah, C., et al.: The building blocks of interpretability. Distill **3**(3), e10 (2018)
23. Osherson, D.N., Stern, J., Wilkie, O., Stob, M., Smith, E.E.: Default probability. Cogn. Sci. **15**, 251–269 (1991)
24. Park, D.H., et al.: Multimodal explanations: justifying decisions and pointing to the evidence. In: IEEE CVPR (2018)
25. Patterson, G., Xu, C., Su, H., Hays, J.: The sun attribute database: beyond categories for deeper scene understanding. IJCV **108**, 59–81 (2014)
26. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: BMVC (2018)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: ACM SIGKDD, pp. 1135–1144. ACM (2016)
28. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: IEEE ICCV (2017)
29. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. CoRR abs/1312.6034 (2013)
30. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. In: ICLR (2015)
31. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
32. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE CVPR (2015)

33. Tokmakov, P., Wang, Y.X., Hebert, M.: Learning compositional representations for few-shot recognition. arXiv preprint arXiv:1812.09213 (2018)
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
35. Wang, Y., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a CNN for fine-grained recognition. In: IEEE CVPR (2018)
36. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI **41**, 2251–2265 (2018)
37. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: IEEE CVPR (2018)
38. Xu, K., Park, D.H., Yi, C., Sutton, C.: Interpreting deep classifier by visual distillation of dark knowledge. arXiv preprint arXiv:1803.04042 (2018)
39. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
40. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. IJCV **126**, 1084–1102 (2018)
41. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: pose aligned networks for deep attribute modeling. In: IEEE CVPR (2014)
42. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE CVPR (2016)