Explaining Visual Classification using Attributes

Muneeb Ul Hassan, Philippe Mulhem, Denis Pellerin, Georges Quénot

Explaining Visual Classification using Attributes

Muneeb ul Hassan
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
muneeb_hassan@outlook.com

Philippe Mulhem
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Philippe.Mulhem@imag.fr

Denis Pellerin
Univ. Grenoble Alpes, CNRS, GIPSA-Lab, F-38000 Grenoble, France
Denis.Pellerin@gipsa-lab.fr

Georges Quénot
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
Georges.Quenot@imag.fr

***Abstract*—The performance of deep Convolutional Neural Networks (CNN) has been reaching or even exceeding the human level on large number of tasks. Some examples are image classification, Mastering Go game, speech understanding etc. However, their lack of decomposability into intuitive and understandable components make them hard to interpret, i.e. no information is provided about what makes them arrive at their prediction. We propose a technique to interpret CNN classification task and justify the classification result with visual explanation and visual search. The model consists of two sub networks: a deep recurrent neural network for generating textual justification and a deep convolutional network for image analysis. This multimodal approach generates the textual justification about the classification decision. To verify the textual justification, we use the visual search to extract the similar content from the training set. We evaluate our strategy on a novel CUB dataset with the ground-truth attributes. We make use of these attributes to further strengthen the justification by providing the attributes of images.**

***Index Terms*—Explainable AI, Deep Neural Networks, Interpretability**

I. Introduction

Over the last few years, Convolutional Neural Networks (CNNs) enjoyed the attention of the research community due to a tremendous surge in performance. After the work of Krizhevsky et al. [1], CNNs become the first choice of researchers to solve computer vision problems. With the use of CNNs, we see great advancement in computer vision, sometimes even surpassing human abilities. However there is no clear idea why CNNs outperform traditional computer vision techniques. To open the black box of CNNs, researchers have proposed several approaches to understand what a network is learning, but it still proves to be a challenging task.

Due to lack of understanding of CNNs, we can not build trustable systems. To ensure trustability of systems of CNNs, we must define transparent and explainable models which explain their predictions. This transparency is useful at three stages of artificial intelligence [2]. First, when AI is significantly weaker than humans and not yet reliably 'deployable' (e.g. visual question answering [3]), the goal of transparency and explanations is to identify the failure modes [4], [5], thereby helping researchers focus their efforts on the most fruitful research directions [2]. Second, the goal is to establish trust and confidence in users when the AI is reliably deployable. Third, when AI is significantly stronger than humans (e.g. chess or Go [6]), the goal of explanations is in machine teaching [7] i.e., a machine teaching a human about how to make better decisions [2].

Interpretability refers to a technique which produces explainable models while maintaining the prediction accuracy and enable humans to understand and trust the system. Decomposable pipelines where each stage is hand-designed are thought to be more interpretable as each individual component assumes a natural intuitive explanation [2]. Deep Models give great performance but are not interpretable and their decision process is vague and there is no formal way to explain why it reached to the specific decision. Due to lack of proper justification, CNNs are considered as black boxes. In order to open this black box, several approaches are proposed for understanding the behaviour and decision of the network. The goal is to explain the decision of classification decision taken by neural network.

We define here a visual justification system, namely **EVCA** (for Explaining Visual Classification using Attributes), which produces an explanation for the classification of one input image, providing the respective class label and explaining why the predicted label is appropriate for this image. We condition language generation on the features produce by the fine grained classifier. Other captioning methods relies on visual features generated from a common network like VGG16, VGG19, ResNet etc. which are pre-trained on ImageNet. Our model uses fine grained recognition to produce strong image features [8]. The model learns to generate a sequence of words using an LSTM. We also justify the classification by using visual search. By using the image features, we search for the relevant images in the training set and retrieve top K relevant results using pairwise distance as a similarity measure. Our objective behind proposing this approach is two fold: Justify classification decision using the textual sentence and also justify it by retrieving relevant images from the training set. We start by introducing a CNN architecture for fine grained classification followed by RNN architecture and visual search.

In the following, section II presents related works. We then describe the methodology and model proposed in section III, before presenting the experimental set-up in section IV and results in section V obtained. We conclude in section VI.

II. Related Work

The need for explaining and justifying automatically generated predictions has been discussed in various contexts,

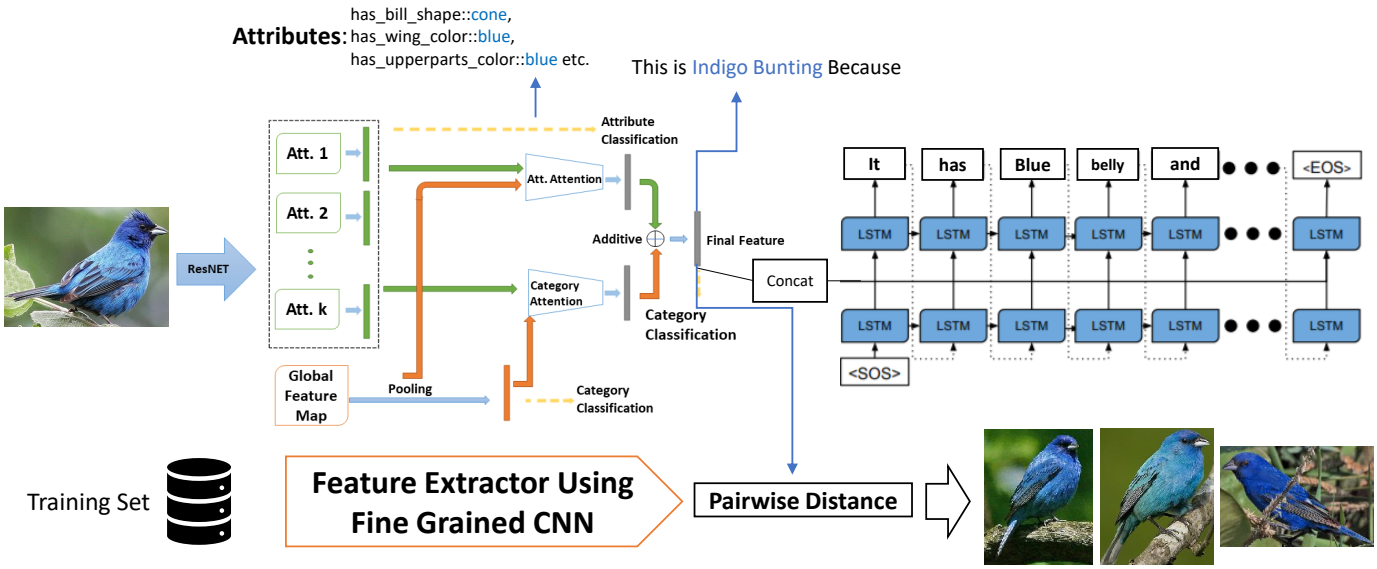

Fig. 1. EVCA generates visual explanation with classification category and also with list of attributes associated with the image. Additionally, it extracts the similar images from the training set using pairwise distance. The images retrieved also contain the attributes. These two parts justify the classification decision.

beginning with experts system in 1970's [9], [10] and it also been studied from a psychological perspective [11] [12]. It is particularly used in high risk application such as medicine where physicians rated the ability to explain decisions as the most highly desirable feature of a decision-assisting system [13]. It is also essential in consumer-facing applications such as Recommender Systems [14], [15] and context-Aware Applications [16], [17].

Explainable AI have been growing rapidly because of the increasing interest in introspective deep neural networks. Many approaches try to explain the decision of deep neural network and to justify the classification result. Zeiler et al. [18] use the deconvolutional approaches to visualise the activations of inner layers of convolutional network, whereas [19] [20] use discriminative patches. [21] proposed automated textual explanation of images for image understanding. Also, LSTM [22] has been used by [23] [24] to generate visual explanations: It uses a loss function based on reinforcement leaning that learns the class specificity to generate sentences. [25] presented a pointing and justification explanation model which provides a joint textual rationale generation and attention visualisation. The model both visually points to the evidence and justifies a model decision with text. [26] [2] uses the heat maps/attention maps for visual explanations by indicating the regions of the image which are most important for the decision.

Explanation Systems can either be introspective systems or justification systems. Introspective systems are designed to reflect the inner working and decision process of deep neural network whereas justification systems are designed to explain which visual evidence supports a decision. [18] [27] [28] define introspective explanations where the model's inner working and decision process are highlighted. Justification explanations models are presented in [23] [29], they use the discriminative image attributes for reasoning process. [25] argued that both systems are useful although justification systems are not necessarily helpful to an AI researcher to debug AI component. Justification system is core to AI problem and it is an AI challenge to answer, "which species a bird is?" but [25] claims also that it is a fundamental AI challenge to answer, "why would one say this image is related to specific bird species". In the work presented here, we justify the decision of classification task by using textual justification, with visual search and related attributes of object.

III. Methodology

Our model is as shown in Fig. 1 explains how a classification decision is made (i) by generating the textual description and explanation, (ii) by predicting the attributes for the specific class which are also present in the textual description and (iii) by retrieving the similar content from the training set to justify what has triggered the particular decision, e.g., "This is (Object Classified) because (Justification)". As we summarise in Figure 1, our model involves four parts: (1) a category classifier which predicts the class i.e *Indigo Bunting* shown in Fig. 1, which uses the fine grained CNN architecture to extract features from images; (2) a textual explanation generator, which generates textual explanation and description about the image content i.e *It has blue belly and ...*; (3) a visual search, which uses the feature vector from the fine grained classifier and retrieve the top K relevant results from the training set as shown in Fig. 1; (4) a attribute classifier, which gives the attributes present in the textual justification and in images retrieve by visual search like *has_bill_shape::cone, has_wing_colour::blue etc* as shown in Fig. 1. We ensure that the final output of the system fulfil the criteria of justification of CNN Classification decision.

A. Convolutional Feature Encoder

We use a fine grained classifier to encode the visual features. Our CNN is based on Attribute Aware Attention Module [8]. The model learns the discriminative features for fine grained classification. It consists of two branches: an attribute branch, a category branch and attention modules. The model uses attribute information to distinguish different categories such as birds from two different species e.g "indigo bunting" and "Lazuli bunting" that both have "blue head" and "cone shaped beak" but different breast colour. The attention mechanism is used to learn basic representation and the important attribute features used to refine category features for classification [8].

The model is composed of shared CNN, namely ResNet pretrained on ImageNet which extracted high level features. We got the feature map of $2048 \times 7 \times 7$ after omitting the last dense layer which shared by the category branch and attribute branch. The category branch produces the category embedding after a global pooling layer. Similarly, the attribute branch also gets the attribute embedding vector for every attribute.

There are two different attention modules (1) Attribute attention which takes the category feature map and K attribute embedding as input and produces the attention map for regional features [8]. The attribute attention selects the K-th attribute by using the attribute-guided attention weights which are as follows:

$$m^{(k)} = \sigma(V^T a^{(k)}) \tag{1}$$

where $\sigma(x)$ is a sigmoid function. We get K attention maps for all the K attributes. These attention maps are merged via max-pooling to get final attention map. These attention maps are then multiplied with category feature map and summed to produce category representation $f^{(region)}$,

$$f^{(region)} = \frac{1}{L} V m^{(region)} \tag{2}$$

Similarly, (2) Category attention takes K attributes embedding and the category embedding and compute the weights similarly to attribute attention.

$$s^{(attr)} = \sigma(A^T v^{(category)}) \tag{3}$$

The final feature is computed by adding the weighted category features and weighted attribute features. These features contain both the information about category and attributes, and they are passed to the two stacked LSTM which learns how to generate an explanation conditioned on these features. These features are also used for category classification by applying a softmax function.

B. Recurrent Neural Network

The features provided by the CNN are now passed to the two stacked LSTM which generates the sentence conditioned on visual features. The first LSTM receives the previously generated word w_{t-1} as input (at time $t = 0$ the model receives a "Start-Of-Sentence" token) and produces the output l_t. The second LSTM receives the image features generated by our fine grained CNN and the output of first LSTM and outputs the probability distribution $p(w_t)$ over the next word. The word w_t is generated by sampling from $p(w_t)$ at each time step. Generation continues until an "End-Of-Sentence" token is generated [23].

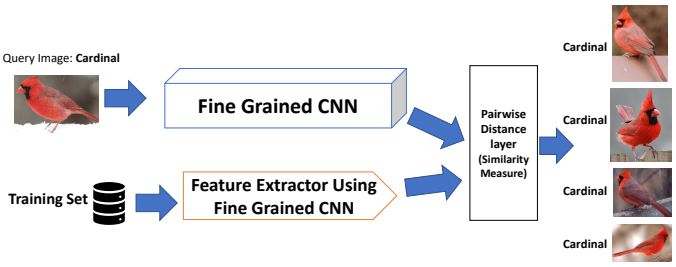

Fig. 2. The Visual Search uses the Fine Grained CNN to extract image features and to compute pairwise distances with images of the training set.

C. Visual Search

In order to achieve justifiable results we take inspiration from human vision. We exhibit images related to the input image retrieved from the training set to achieve our goal. Visual Search uses an image as a query and tries to retrieve the similar object from the training set. The fine grained CNN is used to extract the features from the input image and retrieves the top K relevant results using pairwise distance as a similarity measure. The Pairwise Distance pd is a classical P-norm distance and computed as follows:

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4}$$

where x represents the n-dimensional query image vector and y represents the feature vector of an image from training set. If high distance is observed that means images under observation are dissimilar and if small distance is observed then the images under observation are highly likely have similar context.

IV. Experimental Setup

A. Dataset

In our study, we used well-known Caltech UCSD Birds 200-2011 (CUB) dataset [30]. CUB dataset contains 200 classes of North American birds species and 11,788 images in total. The dataset also comes with attributes for every bird. There are total of 312 attributes for every category of attributes like, bill shape, bill colour, bill length, eye colour etc. Some examples of attributes are given below:

- has_bill_shape::curved_(up_or_down)
- has_bill_shape::cone
- has_bill_shape::hooked
- has_wing_colour::blue
- has_wing_colour::brown
- has_wing_colour::iridescent
- has_wing_colour::purple

We choose CUB dataset because it provides the attributes of every bird class and also there is an extension of dataset

Fig. 3. Visual Explanation Generated by the EVCA justification system where attributes are verified by ground-truth and predicted attributes and these attributes can be find in the images extracted from training set.

which has been done by [31], where they collected 5 sentences for each image. These sentences describe the content of the image, e.g., "This is a bird" but also give detailed description of the bird by mentioning their attributes e.g., "it has cone shaped beak, red body and a grey wing." We selected this image-sentence dataset because every image is belong to a certain class and therefore sentences and as well as images are associated with a single label. The sentence also contains the features of the bird present in the image which make this dataset unique for the visual justification task. The sentence collected in [31] were not collected for the visual explanation task that is why it does not describe why the image belongs to certain class but a descriptive detail about each bird class [23].

B. Implementation

The image features are collected from the penultimate layer of the fine grained CNN. One hot vectors are used to represent input words at each time step and learn a 1000-dimensional embedding before inputting each word into the 1000-dimensional LSTM. We use TensorFlow [32] for our experiments. We reported all the results using CUB standard test set. We train our model with batch size of 64 for 150 epochs. Adam is used as an optimiser with cross-entropy loss. The starting learning rate was 0.001. The Euclidean distance is used to compare pairs of images, so $p = 1$ in equation 4.

V. Results

To justify a classification result, we generate the text from our model with category label and attributes labels. Furthermore, we demonstrate the justification by retrieving the similar images from the training set.

Fig. 3 shows some examples of the our justification system. The EVCA justification system predicts the class label ("Wilson Warbler, American GoldFinch, Florida Jay") and then the justification conjunction ("because") is followed by a textual justification of the classification decision produced by the model.

The first example in Fig. 3 is of Wilson Warbler, where our justification system specifies that the Wilson warbler contains a yellow belly and a yellow breast. We justify this decision by looking at the ground-truth attributes and also the attributes predicted by our justification system. The generated sentences contain the attributes essential to the specific image. We also justify the classification decision by exploiting training set. The images retrieved from the training set for a particular bird class also strengthen the understanding of why a particular image is classified into a particular category. The right of Fig. 3 presents 3 images from the training set that belong to the same class. Similarly, for second and third examples of Fig. 3, where the textual justification contains the attributes present in the query image, we see it from the predicted attributes and the images retrieved from the the training set the prediction is correct.

Despite our efforts, all the attributes are not always present correctly. In Fig. 4, let us focus on the first example with a query image of a "Common Raven": the textual justification mentions one incorrect attribute which is "long neck", and wrong images are extracted from the training set. To explain this, we see that "Common Raven", "Fish Crow", "American Crow" and "Common Crow" are all black, which makes these classes hard to distinguish. Similarly, for the second example, where the classifier predicted "white necked raven", the textual justification only predicts the correct bird colour but does not mention the nape colour (which is white): it mistaken the nape with the chest. This is wrong as White necked Raven

Fig. 4. Some negative examples predicted by the EVCA justification system, where it is able to predict some attributes but those are also common in other classes.

is specified by the white colour on its nape. Similarly, the images extracted from training set do not justifying correctly the decision.

TABLE I
COMPARISON OF EVCA WITH BASELINE MODELS OF [23]

	METEOR	CIDEr
Definition [23]	27.9	43.8
Description [23]	27.7	42.0
Explanation-Label [23]	28.1	44.7
EVCA	28.2	45.3
Explanation [23]	29.2	56.7

We compare our system with the baseline models of [23] where they reported METEOR [33] and CIDEr [34] score. In table I, [23] trained definition model to generate sentence using only image label as input and Description model is equivalent to LRCN [31], except the features used are from fine grained classifier. Explanation-label is equivalent to Description but in addition it also conditioned on class predictions and Explanation model depends on class condition and uses reinforcement loss. Our result are below the *Explanation* model of [23] (last line of table I) but we do provide additional details for justification like attributes which are present in the sentence and the similar images from the training set whereas [23] only provides the textual justification. We present in table II the Bleu [35] and Rouge [36] scores, in a way to show the interest of our EVCA model.

TABLE II
EVALUATION OF EVCA WITH DIFFERENT EVALUATION METRICS

	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE
EVCA	62.6	54.5	35.5	27.3	45.9

VI. CONCLUSION

In this work, we presented an approach for both experts and non-experts to justify the classification decision of Convolutional Neural Network. For experts, we generate visual justification which contain attributes of the bird present in the image and these attributes also predicted by classification model. For non-experts, we exhibit relevant images to the input image, retrieved from the training set, as an additional information to support the classification decision of CNN. This additional visual information provides non-experts a naive sense to trust on the system. Our proposal was tested on the CUB data set of birds images, and compared to other state of the art approaches, on classical evaluation measures. The results obtained outperform existing comparable works. We also provide additional information like attributes and similar images from training set, which makes it unique. We obtain though some false results which was mainly due to ambiguous appearance of birds in an image or very similar birds classes. Exhibiting false results challenges the classification decision of Convolutional Neural Networks. Our results show why classification decision of Convolutional Neural Networks was wrong and helps non-experts to better understand the final decision. Our proposal provides enough visually perceivable justification to convince both expert and non-experts to trust the classification decision of Convolutional Neural Network.

There are many ways to improve the textual justification for a classification task. For instance, [23] uses reinforcement loss and class labels to generate sentences which focuses on the discriminative properties of visible object. We can incorporate this loss to further improve the system.

ACKNOWLEDGMENT

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) in the context of the DeCoRe project.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[3] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[4] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," *arXiv preprint arXiv:1606.07356*, 2016.

[5] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *European conference on computer vision*. Springer, 2012, pp. 340–353.

[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[7] E. Johns, O. Mac Aodha, and G. J. Brostow, "Becoming the expert-interactive multi-class machine teaching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2616–2624.

[8] K. Han, J. Guo, C. Zhang, and M. Zhu, "Attribute-aware attention model for fine-grained representation learning," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 2040–2048.

[9] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.

[10] W. R. Swartout, "Xplain: A system for creating and explaining expert consulting programs," *Artificial intelligence*, vol. 21, no. 3, pp. 285–325, 1983.

[11] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[12] ——, "Explanation and abductive inference," *Oxford handbook of thinking and reasoning*, pp. 260–276, 2012.

[13] R. L. Teach and E. H. Shortliffe, "An analysis of physician attitudes regarding computer-based clinical consultation systems," *Computers and Biomedical Research*, vol. 14, no. 6, pp. 542–558, 1981.

[14] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000, pp. 241–250.

[15] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," in *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 2007, pp. 801–810.

[16] J. Vermeulen, "Improving intelligibility and control in ubicomp," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010.

[17] B. Y. Lim and A. K. Dey, "Toolkit to support intelligibility in context-aware applications," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 13–22.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[19] T. Berg and P. N. Belhumeur, "How do you tell a blackbird from a crow?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 9–16.

[20] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *Communications of the ACM*, vol. 58, no. 12, pp. 103–110, 2015.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[22] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[23] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[24] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 264–279.

[25] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8779–8788.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[27] V. Escorcia, J. Carlos Niebles, and B. Ghanem, "On the relationship between visual attributes and convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1256–1264.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[29] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 563–578.

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[31] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.