Are These Birds Similar: Learning Branched Networks for Fine-grained Representations

Shah Nawaz[1], Alessandro Calefati[1], Moreno Caraffini[1], Nicola Landro[1], and Ignazio Gallo[1]

[1] Department of Theoretical and Applied Science, University of Insubria, Varese, Italy

***Abstract*—Fine-grained image classification is a challenging task due to the presence of hierarchical coarse-to-fine-grained distribution in the dataset. Generally, parts are used to discriminate various objects in fine-grained datasets, however, not all parts are beneficial and indispensable. In recent years, natural language descriptions are used to obtain information on discriminative parts of the object. This paper leverages on natural language description and proposes a strategy for learning the joint representation of natural language description and images using a two-branch network with multiple layers to improve the fine-grained classification task. Extensive experiments show that our approach gains significant improvements in accuracy for the fine-grained image classification task. Furthermore, our method achieves new state-of-the-art results on the CUB-200-2011 dataset.**

***Index Terms*—Multimodal representation, Two branch network, Fine-grained image classification**

I. Introduction

Fine-grained image classification focuses on discriminating between hard-to-distinguish categories or classes, such as birds [1], [2], flowers [3] and cars [4]. Typically, fine-grained datasets are characterized by large intra-class variance and small inter-class variance which makes the task more challenging, as shown in Fig. 1. In fine-grained classification, categories are generally distinguished by subtle and local differences, such as the color of the belly, the shape of the toe and the texture of feather for the bird. These local features are located at the regions of the object or its parts, therefore it is advantageous to localize the object and its relevant parts. Existing methods employ two-stage learning strategies with the first stage localizing the object or its parts while the second stage extracts the representation of the object or its parts through Convolutional Neural Network (CNN). Features extracted are then used to train a classifier for the final classification task. Despite achieving good results, these strategies suffer from two limitations. First, the discriminative parts play a significant contribution in improving the classification accuracy of fine-grained classification tasks, however, not all parts are useful and contribute to the correct prediction. For example, Huan *et al.* [5] extracted 8 to 15 parts, while Zhang *et al.* [6] achieved better performance with only 6 parts. In addition, He *et al.* [7] pointed out that the best number of parts is an empirical value. The other limitation is that the part locations of the object could not provide information on which part is discriminative from all the others between seemingly similar categories. Considering these drawbacks, we need a new way to assign a value to attributes along with the importance of the parts in which parts are located to improve the fine-grained image classification accuracy. Interestingly, natural language description provide useful information which can be exploited for fine-grained image classification.

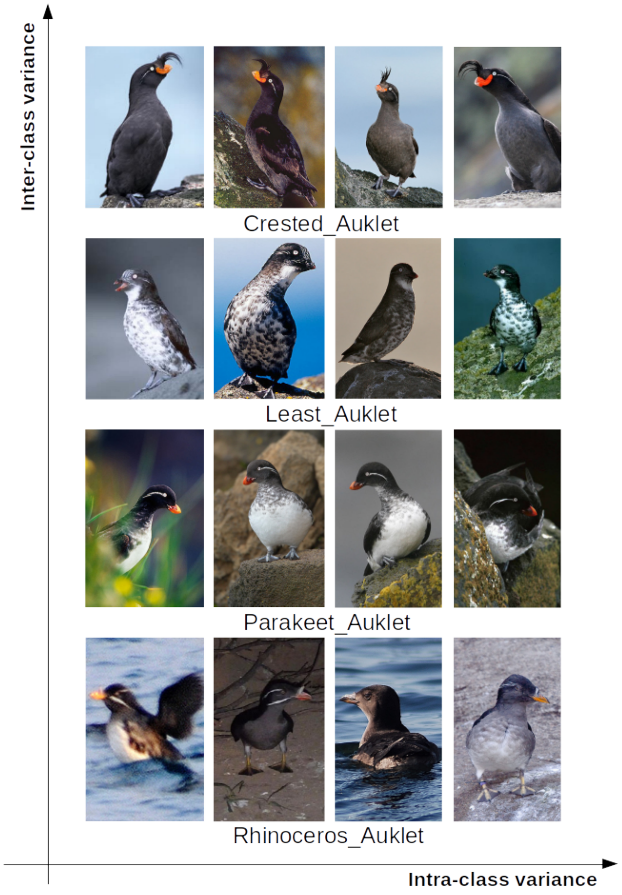


Fig. 1. Images extracted from CUB-200-2011. These examples indicate large intra-class variance and small inter-class variance, which makes the image classification task highly challenging.

978-1-7281-4187-9/19/$31.0©2019 IEEE

In addition, it is well-known that data obtained from images and natural language description may provide enriched information to capture a particular "concept" than either from image or natural language description [8]. The combination of various media types such as natural language description and images is referred as multimodal.

Multimodal data may provide more informative content, however, the challenge is to deal with heterogeneous media types. In fact, each media type has its own representation and structure. For example, natural language is typically represented as discrete sparse word count vectors whereas an image is represented using dense and real-value features [8]. Inspired by the recent successes of multimodal strategies, this paper proposes a Two Branch Network (TBN) combining vision and language for learning joint representation. The vision branch representation is extracted from Navigator-Teacher-Scrutinizer Network (NTS-Net) [9], which localize interesting and discriminative regions without the need of bounding-box or part annotations. The language branch representation is obtained through Bidirectional Encoder Representations from Transformers (BERT) model [10]. Finally, representations are fused after passing through a Multi-Layer Perceptron (MLP). The proposed approach is evaluated on a highly challenging Caltech-UCSD Birds (CUB-200-2011) dasetset [11]. Our approach obtained state-of-the-art performance on CUB-200-2011 dataset.

The rest of the paper is organized as follows: we explore the related literature in Section II; details of the proposed approach are discussed in Section III, followed by experiments in Section IV. Finally, conclusions are discussed in Section V.

II. Related Work

In this section, we summarize relevant background on previous works on fine-grained image classification and multimodal fine-grained classification.

A. Fine-grained Image Classification

Fine-grained image classification has moved from multistage strategies based on traditional hand-crafted features [12], [13] to multistage approaches based on Convolutional Neural Network (CNN) [14], [15]. More recently, end-to-end CNN approaches consisting of unified localization and classification have replaced multistage fine-grained image classification [16], [17]. The vision stream of our proposed approach is extracted from NTS-Net scheme [9], which obtained discriminative parts of the objects.

B. Multimodal Fine-grained Classification

In recent years, several strategies have been proposed in the field of multimodal representation and learning. Although each task is different from others, the underlying principle is relatively the same: to achieve semantic multimodal representation. Recent years have seen a surge in tasks based on multimodal data including classification [18]–[20], Visual Question Answering [21], [22], multimodal named entity recognition [23], [24], cross-modal retrieval [25], [26] and verification [27], [28]. Typically, exiting multimodal strategies employed separate networks to extract representations of each media types and a supervision signal is configured to bridge the gap between these media types. He *et al.* [7] brought the multimodal representation and learning into fine-grained image classification to boost the performance. The approach utilized natural language description to identify the discriminative parts in the associated image. Wang *et al.* [25] presented cross-modal approach with two-branch network consisting of two layers of nonlinearities on top of image and text representations. Similarly, we presented a two-branch network built on top of the image and text representation for the fine-grained image classification task. Though our approach is similar in nature to the one presented by Wang *et al.*, however, there is a fundamental difference. Wang *et al.* employed a two-branch network to solve the cross-modal retrieval task while our approach present a strategy to tackle fine-grained image classification.

III. Proposed Model

The proposed approach leverages on a multimodal concept: to provide the classification of a sample, we exploit text descriptions written in natural language jointly with visual information coming from the image. In the initial stage the pipeline is divided into two streams running in parallel.

A. Visual Stream

The first stream, called Visual Stream, deals with images and converts them as real-valued vectors. Let $I \in \mathbb{R}^{d\times d}$ be an image of the dataset, it is converted into a vector $f_i \in \mathbb{R}^k$ through a NTS-NET [9] model, pre-trained on CUB-200-2011 dataset which extract features. Formally:

$$\phi_v : \mathbb{R}^{d\times d} \rightarrow \mathbb{R}^k \tag{1}$$

with $k = 200$.

B. Text Stream

The second stream, called Language Stream, uses the BERT model proposed in [10]. Given a dictionary $D = \{w_1, w_2, \dots, w_n\}$ where $|D| = n$, we convert a text T into a vector $t = [v_1, v_2, \dots, v_n]$ with the same length of the dictionary. Then we applied the BERT model which converts a text into a feature vector. Formally:

$$\phi_t : \mathbb{R}^d \rightarrow \mathbb{R}^k \tag{2}$$

with $k = 200$.

Both streams are used as feature extractors to train a Multi-Layer Perceptron (MLP) with a Softmax as activation function. The best MLP configuration found experimentally contains 200 neurons in input and output with only one hidden layer with 1000 neurons. The two output layers are summed before being passed to the softmax. No activation functions were used for the two MLPs, so activations are implemented only by the weighted sum of the input plus bias. Each feature extractor gives in output real-valued vectors with a size of 200. Thus,

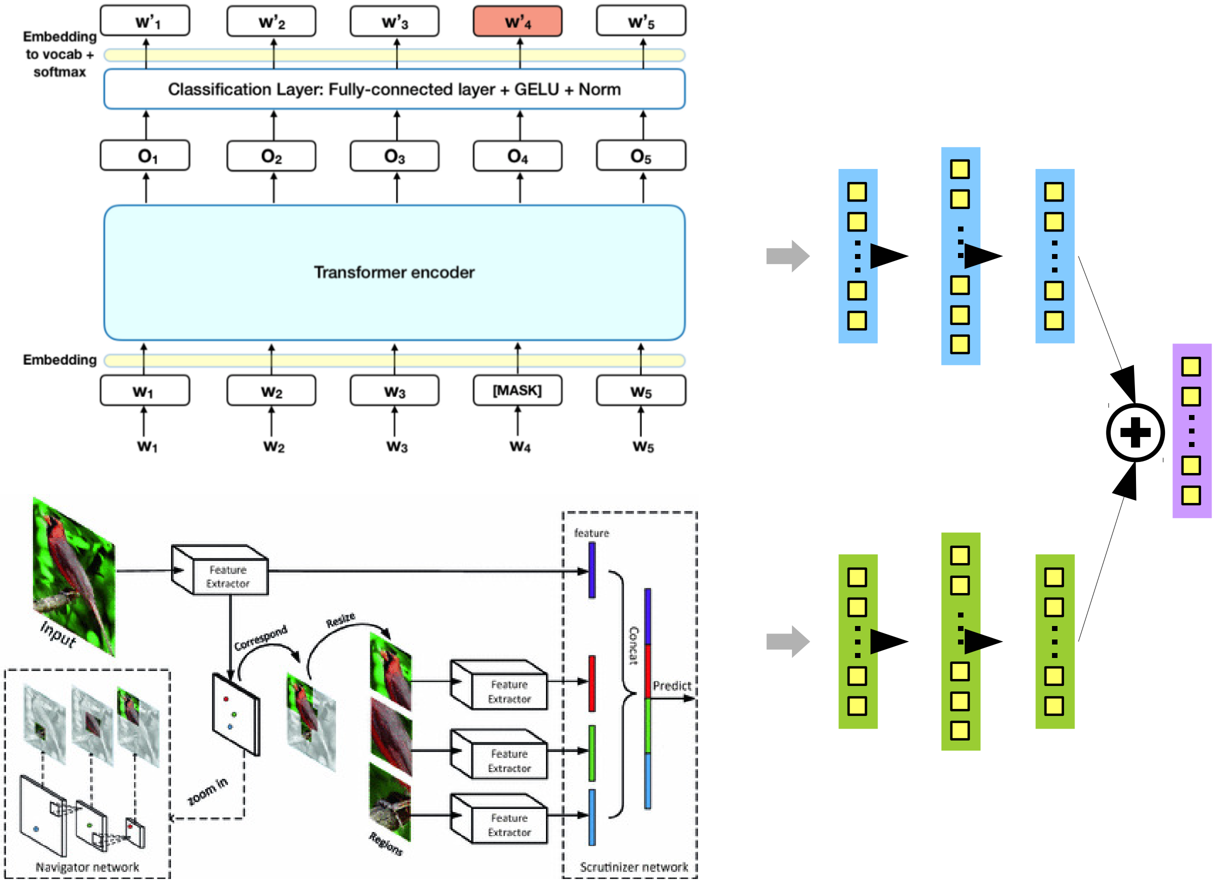


Fig. 2. The proposed approach extracts image and text representation from NTS-NET [9] and BERT [10] models. Finally, representations is fused after passing through a MLP.

the output is obtained by applying the softmax to the last fully connected layers x_1 and x_2 of the two MLPs. More formally:

$$o = softmax(W(x_1 + x_2) + b) \tag{3}$$

Fig. 2 shows details of the proposed model along with pretrained NTS-Net and BERT models.

IV. Experiments

The implementation details, dataset and experimental setup are explained below.

A. Implementation Details

The NTS-NET [9] uses a Navigator Network trained from scratch with $M = 6$, where M denotes the number of regions, $K = 4$, where K means that the top K region will be selected and image size of 448×448 pixels, It uses also a pre-trained ResNet-50 [29] model as feature extractor with Batch Normalization for regularization, Momentum SGD with initial learning rate 0.001 and after 60 epochs it is multiplied by 0.1 and weight decay value as 0.0001.

Pre-trained BERT [10] (base version) for text processing is used, using learning rate value of 0.00002 and weight decay value as 0.00001.

B. Dataset

All experiments have been performed on the challenging fine-grained image classification benchmark dataset CUB-200-2011 [11]. It contains 11788 images of 200 species of birds split in 5994 and 5794 images for train and test respectively. Each image contains:

- 15 part locations
- 312 binary attributes
- 1 bounding box containing the bird within the image

In addition to these information each image is accompanied by 10 fine-grained sentences. Some random examples extracted from the dataset is show in Fig. 3. These natural language descriptions are collected through the Amazon Mechanical Turk (AMT) platform. Descriptions represent features of the bird without including any information about background and actions. Fig. 3 shows characteristics of the dataset. For each line there are 3 samples extracted from a specific class. As it can be seen parts are described with natural language descriptions.

C. Experimental Setup

Four experimental settings have been performed. To evaluate the effectiveness of our approach, we compared the

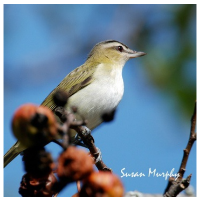
Class	Vision	Natural Language Description
Green Jay		– This bird has a black pointed bill, with a black and green breast. – A colorful bird with shades of green, blue, and black. the beak is thick and pointed. – Bird has a bluish cheek patch and a black eyebrow with a bluish spot. . . .
Common Yellowthroat		– This bird has a brown crown as well as a yellow throat. – This yellow breasted bird has a contrasting brown back and white flanks and thighs. – A small multi colored bird with a white breast and yellow back feathers. . . .
Red Eyed Vireo		– This beautiful gold and gray colored bird had a sharp pointed beak and black tail. – This has a white underbelly with yellow and gray short feathers and a striped face. – A bird with a white belly and a dark brown eyebrow. . . .

Fig. 3. Some random image-text pairs taken from the CUB-200-2011 dataset.

single modal pipelines against the one which merges textual information with visual ones.

- **Textual-stream**: In this setting, we used only text descriptions and classified it with the BERT model.
- **Visual-stream**: In this setting, we exploited the visual information with the NTS-NET to extract features and perform the classification task.
- **Proposed Two Branch Network**: In this experiment, we fused text and visual information with MLPs to perform multimodal classification.
- **Parameters**: In this experiment, we explored various parameters of our model in terms of the topology of the fully-connected layers.

Results of first three experiments are summarized in the Table I as classification accuracy. It is clear from results that the proposed Two Branch Network outperforms accuracy values of each branch. In addition, this result reaffirm well-know observation from existing works where multimodal performs better than individual modalities [18].

To understand which was the best configuration of the proposed model we tried different configurations of the output layers used to combine text and visual streams. In this case we only used the training set of our dataset, dividing it into training and validation. To merge the two streams we used two MLPs with different number of layers and neurons. In particular we used two MLPs configured in the same way for each experiment and in each experiment we changed

- number of hidden layers: 1, 2 and 3
- number of neurons for hidden layers: 500, 1000, 2000

The best result was obtained using two MLPs with only one hidden layer with 1000 neurons.

Method	Accuracy (%)
Visual-stream [9]	87.50
Textual-stream [10]	65.00
Two Branch Network (Proposed) (Textual+Visual streams)	**96.81**

TABLE I
COMPUTED ACCURACY FOR DIFFERENT USED METHODS.

D. Comparison with state-of-art methods

For comparison purpose, we adopt 7 state-of-art fine-grained image classification strategies. Table II shows the comparison results on CUB-200-2011. In our experiments, our approach obtains near 10% higher accuracy than the best performing result of VTRL [30]. The boost in performance is due to the fact that discriminative parts obtained from NTS-NET model is further enhanced with natural language descriptions.

V. CONCLUSION

In this paper we tackled the fine-grained classification task through the Two Branch Network to exploit combined visual and textual information. The proposed approach leverages on two state-of-the-art models for feature extraction: Navigator-Teacher-Scrutinizer Network for images and a Bidirectional Encoder Representations from Transformers model for the text representation. Results achieved show that the approach is effective and boosts significantly the accuracy obtained with existing methods with an improvement of around 10%.

Method	Accuracy (%)
TBN (Proposed Approach)	**96.81**
VTRL [30]	86.31
CVL [7]	85.55
PD [6]	84.54
Spatial Transformer [31]	84.10
Bilinear-CNN [32]	84.10
NAc [33]	81.01

TABLE II
Comparisons with state-of-the-art methods on CUB-200-2011, sorted by amount of annotation used. Our Approach indicates our full method combining vision and language.

References

[1] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2011–2018.

[2] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," *California Institute of Technology*, 2011.

[3] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 811–818.

[4] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[5] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked cnn for fine-grained visual categorization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1173–1182.

[6] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1134–1142.

[7] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5994–6002.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[9] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 420–435.

[10] C. Alberti, K. Lee, and M. Collins, "A bert baseline for the natural questions," *arXiv preprint arXiv:1901.08634*, 2019.

[11] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," no. CNS-TR-2011-001, 2011.

[12] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 729–736.

[13] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3466–3473.

[14] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5546–5555.

[15] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 842–850.

[16] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1143–1152.

[17] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[18] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[19] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5198–5204, 2018.

[20] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 5. IEEE, 2017, pp. 36–41.

[21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *Proceedings of Empirical Methods in Natural Language Processing, EMNLP 2016*, pp. 457–468, 2016.

[22] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, vol. 3, no. 5, 2018, p. 6.

[23] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] O. Arshad, I. Gallo, S. Nawaz, and A. Calefati, "Aiding intra-text representations with visual context for multimodal named entity recognition," in *International Conference on Document Analysis and Recognition*, 2019.

[25] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.

[26] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[27] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[28] A. Nagrani, S. Albanie, and A. . Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88.

[29] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[30] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 1452–1457.

[31] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[32] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[33] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1143–1151.