ADVERSARIAL MULTI-LABEL PREDICTION FOR SPOKEN AND VISUAL SIGNAL TAGGING

Yue Deng[1], *KaWai Chen*[1,2], *Yilin Shen*[1], *Hongxia Jin*[1]

[1]AI Center, Samsung Research America, Mountain View, CA, USA
[2]University of California, San Diego, La Jolla, CA, USA

ABSTRACT

We introduce an adversarial multi-label classification (ADMLC) framework to improve the robustness and performance of existing algorithms on multi-domain signals. The core contribution of our ADMLC is the innovation of an 'adversarial module' that serves as a critic to provide augmenting information to improve supervised learning in multi label classification (MLC) tasks. Our approach is not intended to be regarded as an emerging competitor for many well-established algorithms in the field. In fact, many existing deep and shallow architectures can all be adopted as building blocks integrated in the ADMLC framework. We show the performance and generalization ability of ADMLC on diverse tasks including audio and image tagging.

Index Terms— multi label prediction, audio tagging, adversarial learning

1. INTRODUCTION

Learning to predict multiple labels/attributes $y \in \mathbb{R}^n$ from raw data $x \in \mathbb{R}^m$ is a fundamental topic in machine learning which has inspired a quite large number of real world applications[1]. Unlike traditional multi-class classification with a one-hot label vector y (i.e. $\sum_j^n y[j] = 1$ [1]), multi-label classification (MLC) allows the input data x to be associated with multiple classes (i.e.$\sum_j^n y[j] \geq 1$). Multi-label classification is much more difficult than multi-class classification because the former inevitably involves more complicated label structures than the latter. In audio signal processing society, MLC has encouraged a wide range of real world problems including audio tagging, spoken documents summarization and music emotion detection. A benchmark deep learning solution for audio-based MLC has been proposed in [2] and are then further enhanced by the attention model [3]. In the general machine learning filed, prevalent MLC methods could be profiled as two types: shallow and deep approaches.

Among shallow MLC approaches, label embedding is a representative one that projects both input data and its corresponding multi-label vector to a latent space, e.g. maximizing their correlations in CCA space [4]. Alternative label embedding approaches include principal label space dimension reduction (PLST) [5] and sparse local embedding (SLEEC) [6]. PLST and its conditioned version (CPLST) seek the label projection by revealing its principle components. SLEEC was inspired by ensemble and sparse learning that is shown to be effective in handling rare labels. Low rank empirical risk minimization for multi-label classification (LEML) [7] considers missing labels could be linearly spanned by known labels and recovers unknown labels by low rank optimization [8]. Unlike these label embedding approaches, label selection directly decreasing the label dimensions by selecting a subset of label columns from the complete set. However, these label selection approaches rely too much on heuristic search [9].

Deep learning allows MLC training in a more favorable end-to-end manner. Back-propagation based multi-label learning (BPMLL) was an early attempt to use neural networks for MLC [1]. More advanced BPMLL implementations were considered in [10] where various MLC losses and tricks were discussed. Deep-CPLST is a deep extension of CPLST proposed in [11]. Recently, the CNN-RNN [12] structure was proposed in the computer vision field. Such approach exploits a CNN to learn image features and a RNN to iteratively generate multi-labels. The Canonical-correlated auto-encoder (C2AE) [11] implements a deep CCA network (encoder) followed by an auto-encoder network (decoder) to translate the joint embedding as the multi-label vector.

However, these off-the-shelf MLC approaches can hardly be directly applied to analyzing some information-rich audio signals due to the following two reasons. The first known challenge is the inability to deal with 'so many' labels [9, 13]. In spoken documents summarization, the inherent document can naturally cover a huge range of concepts and hence lead to thousands of multi-labels for prediction. The second noticeable challenge is about the generalization capability of existing models when they are only trained with limited labeled data. In image domain, large scale training datasets [14] have already been well built and the acquisition ofenoughtraining samples are relatively cheap. However, this scenario does not apply to audio processing in which we can only have access to very limited training samples for spoken data, e.g. in the audio-tagging task[2]. Complicated MLC architectures usu-

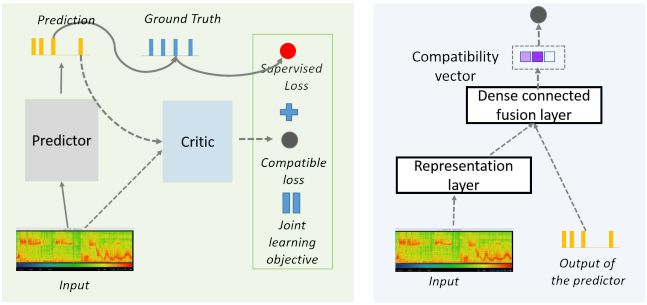
[1] $y[j]$ denotes the jth dimension of the label vector

ally gain the bad reputation of poor generalization. In addition to conventional approaches like regularization and dropout, it is quite plausible if some anti-over-fitting mechanism could be inherently incorporated in the learning paradigm.

In this work, we consider leveraging the power of the breakthrough adversarial learning [15, 16] to improve the performance, robustness and generalizations of existing MLC approaches. The adversarial MLC (ADMLC) framework is implemented with a predictor and a critic module, in which the latter could iteratively criticize the multi-label predictions from the former. These two modules are synchronized with the help of the adversarial loss. It is not hard to conceive that the augmented adversarial critic module can offer auxiliary learning objective in addition to the typical supervised loss in MLC. ADMLC is a general framework that is naturally compatible with a lot of existing MLC algorithms developed in both shallow and deep fields. In conclusion, ADMLC does not play the role of a new competitor but is more preferable to be regarded as a complementary solution that could further enhance existing MLC algorithms.

2. ADMLC MODEL

2.1. Framework



(a) Adversarial MLC framework (b) Architecture of the adversarial critic

Fig. 1. (a) An overview of Adversarial Multi-label classification (ADMLC). The solid gray arrows indicate the 'supervised learning' process. The dotted arrows indicate the 'adversarial critic' process between the 'critic' and the 'predictor'. The critic compared predicted multi-label vector with the raw input and intends to assign a low score for the predicted label but granting a high score for the natural pair of input data and ground-truth multi-label vector. The aforementioned 'adversarial learning' process leads to an minmax optimization involving in a gambling process. (b) The detailed configuration of the 'critic' module in (a).

The intuition of ADMLC learning could be conceptually comprehended by imaging the interaction between a student and a professor in a quiz process. Initially, a student takes a quiz and turns his answer to the professor. The professor criticizes on the student's answer by finding out all his mistakes. In our ADMLC learning framework, we design two modules to respectively mimic these two roles: predictor (student) and critic (professor) in Fig 1 (a). In the aforementioned process, we have also observed two types of gambling interactions among these two roles. On the student's side, he/she expects to get a high score on his/her quiz from the professor. However, the professor should play a very strict role to identify all his/her mistakes and make a fair score. These two types of gambling interactions yield the problem to the typical paradigm of adversarial learning.

Mathematically, we use $f_P(\cdot)$ and $f_C(\cdot)$ to respectively define the transformations of the 'Predictor' and the 'Critic' modules in Fig.1(a). Both these functions can be implemented with neural networks. Initially, the predictor inputs the raw data x_i and predicts its corresponding multi-label vector $\hat{y}_i$. Then, $\hat{y}_i$ is examined by two learning objectives, i.e. the supervised loss and the compatible loss.

First, we compare $\hat{y}_i$ with the ground truth label y_i to define its 'supervised losses':

$$L_P(\hat{y}, y) = dist(\hat{y}, y) \tag{1}$$

where $dist(\cdot, \cdot)$ defines the distances/losses of the predicted value and the ground truth. In MLC, l_p losses are not traditional selections due to the weak interpretations of the coherent label structures. In a recent work [17], two advanced losses were considered for MLC problems including pairwise ranking (PR) loss and the weighted approximating ranking (WARP). PR loss borrows the concept of hinge loss to penalize the margin between positive and negative labels. WARP is a weighted version of PR loss which further takes the importance of different classes sizes into consideration. In this paper, we adopt the more advanced WARP as the supervised loss.

In addition, we further pass the predicted multi-label vector $\hat{y}_i$ and input data x_i through the critic module to get its 'compatible loss',

$$s_i^{(p)} = f_C(x_i, \hat{y}_i) = sigmoid(g(h_{(x_i, \hat{y}_i)})) \tag{2}$$

where $s^{(p)}$ is the score measuring the compatibility of the input data and the predicted multi-label vector. To better explain the scoring mechanism, we explicitly provide the architecture of the critic in Fig.1(b). The critic module embraces both the raw input and predictor output as dual inputs. The raw input (e.g. a raw audio segment) is first processed by the representation layer (CNN or RNN) and is converted as a compact low dimensional feature vector. This feature vector is fused with the output of the predictor as the joint input to a dense connected fusion neural network [18]. We write out the last three transformations in the critic neural network. In (2), $h_{(x_i, \hat{y}_i)}$ is obtained by fusing the information from both the predicted multi-label ($\hat{y}_i$) and raw data input (x_i); $g(\cdot)$ is fully connected with $h_{(x_i, \hat{y}_i)}$ and outputs a single value; and the sigmoid function transforms the regressed value in the range of $[0, 1)$ as the final score.

2.2. Objectives

After understanding the transformations in these two modules, we will define the optimization objectives for the whole ADMLC framework. The joint learning loss is defined by the following two additive terms balanced by a hyper-parameter λ:

$$\begin{aligned} L_C &= L_P(\hat{y}, y) + \lambda L_{CE}(s^{(p)}, \mathbf{1}) \\ &= L_P(\hat{y}, y) - \lambda \tfrac{1}{N} \textstyle\sum_{i=1}^{N} s_i^{(p)} \log s_i^{(p)} \end{aligned} \tag{3}$$

where the first term L_P is the prediction error (denoted by the red dot in Fig.1(a)), which is the same as the traditional supervised loss counting the differences between the prediction and the ground truth. The second term in (3) is defined from the aspect of adversarial learning. It encourages the predictor to make a prediction that could achieve a high score $s^{(p)}$ after passing through the critic. We follow the same implementation in GAN [15, 19] to adopt a cross-entropy term $L_{CE}(s^{(p)}, \mathbf{1})$ to encourage the predicted labels' scores as high as one. $\mathbf{1}$ is a vector with all entries equal to one. From the critic's point of view, $s_i^{(p)} = 1$ means the predicted multi-label vector $\hat{y}_i$ perfectly matches the raw data x_i and is the highest score could be granted.

On the other hand, the critic learning process is indicated by dotted arrows in Fig.1 (a). In detail, there are two kinds of data to be criticized by the critic network, i.e. the predicted multi-label vector (indicated by solid green arrows) and ground truth multi-label vector (indicated by the dotted green arrows). The purpose of the critic network is to assign a score for the prediction to measure how well the multi-label vector matches the corresponding input data. Accordingly, the raw input data should also be inputted to the critic network. By comparing the multi-label vectors (both predicted and ground truth) with the input data, it intends to give lower scores for those predicted labels while granting high scores for the ground truth labels. Mathematically, a similar cross-entropy term is used here to define the compatible loss:

$$\begin{aligned} L_A &= L_{CE}([s^{(p)}, s^{(g)}], [\mathbf{0}, \mathbf{1}]) \\ &= -\tfrac{1}{N} \textstyle\sum_{i=1}^{N} (1 - s_i^{(p)}) \log(1 - s_i^{(p)}) - \tfrac{1}{N} \sum_{j=1}^{N} s_j^{(g)} \log s_j^{(g)} \end{aligned} \tag{4}$$

where $s^{(p)}$ is the score for predicted labels as defined in (2)th and $s_j^{(g)} = f_P(y_j, x_j)$ is the critic score for the jth ground truth multi-label vector. As reflected in (4), the loss penalize all predicted labels' scores to be very close zero (the lowest score) and encourage $s^{(g)}$ to approximate to one (the highest score). It exactly contradicts to the second term in (3), where the predictor always expects very high scores for its predictions. It hence depicts the adversarial gambling process of these two learning objectives. We should remind that the sample sets for calculating $s^{(p)}$ and $s^{(g)}$ may not be necessarily equivalent. In practical optimization, we will sample a mini-batch of samples $X^{(p)}$ for calculating $s^{(p)}$ but using another randomly sampled set $X^{(g)}$ to calculate $s^{(g)}$. Such random sampling strategy allows more diverse training samples' combinations for this loss and could alleviate over-fitting.

In order to pre-train the critic network, we need both compatible and incompatible pairs. A compatible sample is easily composed by pairing a data point with its ground truth label, *i.e.* (x_i, y_i). We then adopt two straightforward ways to generate incompatible pairs. The first approach is to match a data point with a randomly selected label *i.e.* $(x_i, y_j), j \neq i$. An alternative way is to generate an incompatible multi-label vector $\bar{y}_i$ for a data point x_i by changing or removing some entries on its ground truth label vector y_i. After accumulating both compatible and incompatible pairs, the critic network can be well initialized by regressing all compatible pairs to a score 1 and incompatible pairs to score 0 with a cross entropy loss.

During training, parameters θ_p in predictor network and θ_c in critic network are obtained from Algorithm 1 and can be directly used in inference.During inference, we only pass the testing data x_{test} through the predictor structure in Fig.1 (a) to get the prediction $\hat{y}_{test}$ as the final output vector. The critic network is only used in the training phase but is silenced during inference.

Algorithm 1: ADMLC optimization

Input : A training dataset $\{\mathcal{X}, \mathcal{Y}\}$; the number of recurrent ADMLC unfolding steps T; the steps l for critic network updating;
Initialization : Initialize the predictor and critic.
for *k=1...K* **do**
  /* Predictor learning */
  Sample a minibatch of N samples X and their labels Y from training set $\{\mathcal{X}, \mathcal{Y}\}$;
  Feed-forward X through the predictor and critic in Fig.1(a) to get the multi-output predictions $\hat{Y}$ and critic score $S = f_c(X, \hat{Y})$.
  Use X,Y,$\hat{Y}$and S to calculate the loss in Eq. (3) and back-propogate the loss to update parameters θ_P in predictor network;
  /* Critic learning */
  if *mod(k,l)==0* **then**
    Sample a random batch of M samples $X^{(g)}$ with their ground truth labels $Y^{(g)}$;
    Get Ground Truth evaluation score $G = f_c(X^{(g)}, Y^{(g)})$;
    Take G and S in Eq.(4) for loss calculation and back-propogate the loss to update parameter θ_c in the critic network;
  end
end
Output : Network parameters θ_P and θ_C;

3. EXPERIMENTAL RESULTS

In the experimental part, we will verify the effectiveness of ADMLC on two different domains including document, audio and images. The performance of ADMLC will be compared with other representative methods in these domains,respectively.

3.1. Audio Tagging

Audio tagging is a multi-label classification task that requires to tag all audio event in chunks. We conducted experiments based on a public dataset in the task 4 of the Detection and

Table 1. EER comparisons on seven labels MLC on DCASE 2016 evaluation set. (%)

	c	m	f	v	p	b	o	ave
CE	0.17	0.16	0.18	0.03	0.15	0.00	0.24	0.13
WARP	0.16	0.15	0.18	0.04	0.16	0.02	0.21	0.12
AD-CE	0.17	0.13	0.16	0.04	0.17	0.02	0.20	0.11
AD-WARP	0.16	0.15	0.17	0.04	0.15	0.01	0.21	0.11

Classification of Acoustic Scenes and Events 2016 (DCASE 2016) challenge. The audio recordings in this dataset was accumulated in a domestic environment [20] with seven annotated acoustic event including Broadband noise (b), Child speech (c), female speech (f), male speech (m), other identifiable souds (o), percussive sound (p) and TV or video games(v). These seven events can co-appear in the same audio chunk and thus naturally yield to the MLC problem. The number of recordings is 4387 for the development set and 816 for the evaluation set. Five-fold sets are configured in the development set. We all used these default data splits to perform our consequent experiments and comparisons.

We follow the protocol in [2] to preprocess the raw audio data such as extracting the Mel-Filter banks (MFB) with 40 channels as input feature. Then, we use the same convolution gated recurrent neural network structure (CGRNN) in [2] to implement the predictor module in ADMLC. This because the CGRNN has achieved the better performance than other competing deep neural networks on this dataset. We used both the cross-entropy loss (CE) as in the work [2] and weighted approximating ranking (WARP) loss proposed in [17] as the supervised loss for comparisons purpose. Experimental results on the final evaluation set were provided in Table.1 by using equal error rate (EER) as the accuracy indicator.

From the results, we have observed that WARP loss is a little bit better than the CE loss. Meanwhile, the ADMLC can further enhance the results of these two losses as observed from the second block in Table 1. The best average score was achieved by two ADMLC-based models(AD-CE and AD-WARP). More interestingly, within the ADMLC framework, the performance gap between CE and WARP has become minor.

3.2. Image Tagging

We further consider verifying ADMLC's effectiveness for image attributes prediction. Three image datasets used in this work include the NUS-wide [21], ESP-game [22] and CUB-bird datasets[23]. For ESP-Game, we directly use the provided $1,000$-dimensional bag of SIFT features. For the other two datasets, we extracted $4,096$ image feature from the last layer of a pre-trained VGG-16 network. Images in these three datasets are all annotated with multiple attributes. Image attribute prediction accuracy is conventionally evaluated by micro and macro F1 scores [17]. Micro-F1 combines the precision and recall at the per-class level while macro F1 combining them at the per-sample level.

In this image test, we only consider deep-learning-based MLC methods as building blocks for ADMLC because they are more powerful from previous tests. We follow the same protocol in [11] to randomly select r data points in each dataset for training/validation and the rest are for testing. r is fixed as $15,000$, $18,689$ and $10,000$ on these three datasets, respectively. The random process is repeated for 10 times with the average Micro/Macro-F1 score reported in Table 2. In addition to those deep MLC algorithms used before, we consider two image-based deep learning systems including CNN-RNN [12] and WARP [17] for comparisons. From the result, we have observed that CA2E and BPMLL are better than others. With the ADMLC approach, performances of all deep learning approaches can be further improved. Moreover, the ADMLC's improvements on BPMLL are more significant than on CA2E. BPMLL is designed with a quite simple feed-forward neural network trained by l_2 loss.

Table 2. The performance of ADMLC on image datasets evaluated by micro-F1 and macro-F1 measures

Data		Deep-CPLST		WARP		BPMLL		CA2E	
		Ori.	AD	Ori.	AD	Ori.	AD	Ori.	AD
NUS	MicF1	36.8	39.5	33.5	38.2	38.3	49. 5	41.3	**50.1**
	MacF1	57.2	60.6	53.9	56.7	62.5	67.1	67.6	**68.7**
ESPE	MicF1	5.6	10.1	6.2	9.8	14.8	**15.8**	13.6	14.3
	MacF1	9.7	17.4	6.9	16.3	19.8	**25.3**	22.3	24.2
CUB	MicF1	6.4	9.1	5.8	8.2	7.4	**10.1**	7.8	9.7
	MacF1	14.6	17.1	13.2	15.5	15.8	**20.4**	16.3	20.1

We further report the computational costs and performances of BPMLL and CA2E on this task. For references, the original BPMLL and CA2E methods respectively cost 17s and 41s on the CUB dataset to finish 100 epochs. By adding the critic network, the computational costs increase to 47s and 93s for these two respective methods. All reported time was calculated by running our algorithm in Tensorflow with 8 GPUs.

4. DISCUSSIONS

We proposed the ADMLC framework to improve existing MLC algorithms' performances on difficult tasks. There are still some promising directions deserving our further studies. For instance, critic mechanism is widely used in the reinforcement learning field. We could consider extending our novel concepts of adversarial-critic to solve more challenging reinforcement learning problems. We have also tested ADMLC on other single-output classification and regression tasks. Unfortunately, no significant improvements were observed on those tasks. This limitation may be due to the simplicity of the output structure, where less information could be criticized by the critic.

5. REFERENCES

[1] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[2] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 3461–3466.

[3] X. Yong, K. Qiuqiang, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," *arXiv preprint arXiv:1703.06052*, 2017.

[4] L. Sun, S. Ji, and J. Ye, "Hypergraph spectral learning for multi-label classification," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 668–676.

[5] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[6] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in Neural Information Processing Systems*, 2015, pp. 730–738.

[7] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale multi-label learning with missing labels," in *International Conference on Machine Learning*, 2014, pp. 593–601.

[8] Y. Deng, Q. Dai, R. Liu, Z. Zhang, and S. Hu, "Low-rank structure learning via nonconvex heuristic recovery," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 3, pp. 383–396, 2013.

[9] W. Bi and J. Kwok, "Efficient multi-label classification with many labels," in *International Conference on Machine Learning*, 2013, pp. 405–413.

[10] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, *Large-Scale Multi-label Text Classification — Revisiting Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 437–452. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-44851-9_28

[11] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent spaces for multi-label classification," 2017.

[12] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285–2294.

[13] E. Loza Mencía and J. Fürnkranz, "Efficient pairwise multilabel classification for large-scale problems in the legal domain," *Machine Learning and Knowledge Discovery in Databases*, pp. 50–65, 2008.

[14] 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.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.

[17] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *arXiv preprint arXiv:1312.4894*, 2013.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[19] Y. Deng, Y. Shen, and H. Jin, "Disguise adversarial networks for click-through rate prediction," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 1589–1595.

[20] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[21] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[22] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 319–326.

[23] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.