Learning to Predict Visual Attributes through Human Disagreement

Andrew McIsaac **Leonardo Bertolazzi** **Rossana Cervesato**

February 12, 2023

Abstract

Attribute prediction has made exciting recent progress with the emergence of attribute-centred datasets. However, the reliability of their ground truth labels does not reflect human disagreement. We present a comparative study of the use of soft and hard labels for this supervised multi-label classification problem. Specifically, we compare our baseline with two models that differently incorporate disagreement as information: class-level soft labels from human uncertainty and self-knowledge distillation. The source code can be found at https://github.com/awmcisaac/glp

Keywords attribute prediction · soft-labels · knowledge distillation · VAW dataset

1 Introduction

Modern research in cognitive science and artificial intelligence usually makes use of large datasets annotated with human judgments. The way these annotations are collected can vary: they may involve the judgment of a single expert, or a group of experts or even non-expert workers, depending on the task of interest. Despite these differences, most annotation designs assume that for each element there is a single preconceived interpretation or objective truth, which correspond to the so-called *gold labels*. In contrast, for large-scale annotation projects, it is likely to encounter instances where humans disagree. These disagreements may be due to misunderstandings or poorly specified annotation schemes, but in many cases the interpretation can be inherently ambiguous or unclear. For example, considering the field of computational linguistics, some popular classification tasks involve labeling text based on inherently subjective judgments, such as sentiment analysis or offensive language detection. Indeed, it would be clearly misleading to rely on gold labels for training or evaluating models in such tasks, as it would impose one subjective interpretation over all others. For the same reason, it would also be useful to find effective ways to turn this disagreement into meaningful information that can improve model training and evaluation.

When gold labels are used in a supervised multi-label classification problem, model targets are represented as discrete classes, with correct classes labeled as 1 and others are assigned a value of 0. When this approach is used, the labels are also called *hard labels*. On the other hand, when it is not assumed that there is a single gold label for each input element, uncertainty can be incorporated in model targets using *soft labels*, which consist of probabilities or confidence scores, where the sum of the scores across all classes add up to 1.

To address the problem of learning from human disagreement, we propose a method to combine hard and soft labels for the task of *attribute prediction* on the Visual Attributes in the Wild (VAW) dataset [Pham et al., 2021]. Then, we conduct a comparison between three models: one model trained using hard labels, one model trained using soft labels, and one model trained using self-knowledge distillation from the hard labels model.

Our main contributions are:

- We present a rule-based pipeline to exploit human disagreement in a multi-label classification task with hundreds of labels.
- We define an attribute prediction setup based on the VAW [Pham et al., 2021] dataset, which offers challenging learning conditions like few-instances per object and a long-tailed distribution.
- We provide a comparative analysis of the use of soft labels and hard labels for the task of attribute prediction.

Additional details on the method and results are available in the Supplementary Materials.

2 Related Work

Uma et al. [2022] conducted a survey in which methods for learning from disagreement in annotations are divided into four broad categories: 1. Methods that assume that a gold label exists for every element, and automatically aggregate annotations into one gold label. 2. Methods that assume that a gold label exists for every element, but use disagreement to either eliminate items whose gold label does not appear to be recoverable due to excessive disagreement among coders or to weigh them. 3. Methods that do not necessarily assume that a gold label exists for every item, and assign a score (usually a probability) computed from the crowd annotations to each label (creating soft labels). 4. Methods that do not necessarily assume that a gold label exists for every item, and use a combination of hard labels and soft labels extracted from annotations. These methods typically use gold labels, but supplement them with information from human annotations, e.g., to weigh an item by its estimated difficulty.

Regarding soft labels, their use has been shown to improve the generalization and robustness of models [Pereyra et al., 2017, Müller et al., 2019]. Two widely used methods to automatically construct soft labels are label smoothing [Szegedy et al., 2016] and knowledge distillation [Hinton et al., 2015]. Knowledge distillation is the process of transferring knowledge from one neural network model, the *teacher*, to another, the *student*. Typically the teacher model is pre-trained on a training dataset using hard labels. Its knowledge is then *distilled* to the student model by adding the additional objective of minimizing a divergence metric between its logits and the logits of the teacher[1]. *Self-knowledge distillation* refers to the special case when the student has the same architecture as the teacher. Student models have been shown to surpass performance of the teacher models in a range of tasks [Furlanello et al., 2018] due to the richer output distribution of the teacher model providing additional information about training samples to the teacher model.

However, both label smoothing and knowledge distillation construct soft labels without exploiting the disagreement that can reside in human judgments. Instead, Peterson et al. [2019] provide a method to obtain soft labels directly from human annotation on CIFAR-10. Their method consists of asking a large number of annotators, approximately 50 per image, to label the whole dataset. With the variation in responses obtained involving a high number of annotators, the authors were able to build a probability distribution for each image more informative than with hard labels. As a consequence, models trained using soft labels derived from human disagreement improved generalization and robustness, but were also better in approximating human uncertainty than models built using other soft labeling methods.

The main tasks for which judgment disagreement has been studied are Part-Of-Speech tagging [Plank et al., 2014, Uma et al., 2020], anaphora/co-reference resolution [Poesio et al., 2019], textual entailment [Pavlick and Kwiatkowski, 2019] and sentiment analysis [Kenyon-Dean et al., 2018] for NLP, whereas in CV the main focus has mostly been on image classification [Peterson et al., 2019, Uma et al., 2020, Collins et al., 2022]. To the best of our knowledge, the study of human disagreement in the task of attribute prediction has not yet been explored.

Attribute prediction is a multi-label classification task that consists of predicting the attributes of one or more objects in an image. It is similar to standard image classification but with the main difference that

[1] This is the most typical setting. Knowledge can be distilled from other parts of the teacher network that are not necessarily the output layer.

while object categories are usually mutually exclusive (for example, if an entity is a horse it cannot also be a dog), multiple attributes can be predicted for a single entity at the same time. The main datasets for attribute prediction are: COCO Attributes [Patterson and Hays, 2016], Visual Genome [Krishna et al., 2016] and VAW [Pham et al., 2021].

3 Research Question

The theoretical question we investigate in this report is: *does incorporating human uncertainty into the data improve model performance for multi-label classification problems?* This question follows the line drawn by the work presented in Section 2. Given the benefit to performance and robustness in simpler image classification settings of exploiting human uncertainty, we expect that similar measures may extend well to the multi-label setting.

To test this hypothesis we chose the task of attribute prediction on the VAW dataset. This experimental setting is suited to test the integration of human uncertainty for multiple reasons. Attributes are inherently nuanced in the sense that they can be perceived as more or less typical depending on the object to which they belong and the example (*prototype* using the terminology of Rosch [1973]) a human being has in mind. For example, for the color attribute "red", the usual red of a red wine is different from the typical red of a parrot or that of a brick. Also, continuing with the example of red wine, different people may have different prototypes of red wine from which any red wine may be more or less distant. This makes attribute prediction a suitable task for producing disagreement among human beings.

In addition, the VAW dataset contains complex images with a large number of attributes and object categories, and each object is annotated with several *positive* and *negative* attributes. These characteristics make VAW suitable to be enriched with information derived from human uncertainty on attributes.

4 Dataset

4.1 Data Collection and Statistics

Visual Attributes in the Wild dataset [Pham et al., 2021] proposes a large-scale dataset covering a wider range of attribute and object categories. VAW is constructed with a large vocabulary of 620 attributes, each belonging to one of the 8 parent classes: `action`, `color`, `material`, `other`, `shape`, `size`, `state`, `texture`. Each instance is annotated with positive, negative, and missing attributes.

We use the training, validation and test set as defined within the VAW dataset. The dataset contains 58,565 images for training, 3,317 images for validation, and 10,392 images for testing. However, automated filtering techniques were used to keep the human annotation and training costs feasible, resulting in a smaller number of images per set. While maintaining the proportions of the VAW dataset, we randomly sampled 8800 instances for the training set, 1295 instances for the test set, and 500 instances for the validation set. In addition, we adapt the setup for generalized zero-shot detection as proposed in Bansal et al. [2018] for the attribute detection task. We defined the splits of the test set for known and novel object classes taking approximately 35% of the total number of objects to be novel, resulting in 336 known and 209 novel objects. This setup is critical for testing the generalization capabilities of models at evaluation time.

4.2 Human Annotation and Class-Level Soft Labels

Annotators were provided images together with object instance names (e.g. "man") and their respective positive attributes (e.g. "tall", "wearing shirt"). To rule out any ambiguity each object instance was highlighted with its bounding box. For every attribute belonging to the list of positive attributes of an object, each annotator had to express a positive or a negative judgment about the appropriateness of the attribute with respect to the object.

Table 1: Kappa agreement score for each attribute class

	action	color	material	other	shape	size	state	texture
Kappa score	0.50	0.45	0.42	0.53	0.53	0.38	0.39	0.71

We collected a random sample of 104 object instances with at least two positive attributes. In total, we had 279 attributes[2] and 5 annotators, each of whom annotated all attributes. Since each attribute belongs to one of the eight parent classes, we measured the Cohen's kappa [Cohen, 1960] agreement score among annotations for each class, and we extended the agreement on classes to the attributes that belong to them. Table 1 shows the kappa agreement for each parent class.

Cohen's kappa agreement score among annotations for each class was then used to create class-level soft labels. We re-weight the positive attributes based on the Cohen's kappa score per class, normalizing to the range 0.7-1, and assigning a value of 0.05 to the unlabeled data.

5 Methods

5.1 Problem Formulation

Let $\mathcal{D} = \{I_i, o_i; Y_i\}_{i=1}^{N}$ be a dataset of N training samples, where I_i is an object instance image (cropped using its bounding box), o_i is the category of the object for which we want to predict attributes, and $Y_i = [y_{i,1}, ..., y_{i,C}]$ is its C-class label vector with $y_c \in [0, 1]$ denoting whether attribute c is positive, negative, or uncertain respectively. Our goal is to train a multi-label classifier that, given an input image and the object name, can output the confidence score for all C attribute labels.

5.2 Model Architecture

Fig. 1 shows baseline network architecture. The model resembles the components of the baseline model proposed in the VAW paper Pham et al. [2021]. We can decompose the information flow in the model into three consecutive steps: feature extraction, feature composition, and finally classification. In the next few subsections, we discuss the details of each step.

5.2.1 Feature Extraction

We use the ImageNet-pretrained ResNet-50 [He et al. [2016]] to generate valuable features for the final classification step. It is of utmost importance to extract features that help in predicting attributes accurately. Some of which require an understanding of the whole image while others are intrinsic to the object. Thus, we extract features at three different levels of the backbone network to enrich the feature extraction process and for enhanced prediction of attributes.

Image feature representation: Given an image I of an object o, let $f_{img}(I) \in R^{H\text{x}W\text{x}D}$ be the D-dimensional image feature map with spatial size $H\text{x}W$ extracted using the second, third or penultimate layer of ResNet-50.

5.2.2 Image-Object Feature Composition

We composed the image feature map of penultimate layer of ResNet-50 with object embedding, as outlined in Pham et al. [2021]'s model. We included a simple object-conditional gating mechanism into the model to acts as a filter that only selects attribute features relevant to the object of interest and suppresses incompatible attribute-object pairs (e.g. parked dog).

Let $\phi_o \in R^d$ be the object embedding vector, $f_{comp}(f_{img}(I), \phi_o) \in R^D$ be the composition module that takes in the image feature map and object embedding. We implement f_{comp} with a gating mechanism as

[2] The number of attributes without repetition is 116.

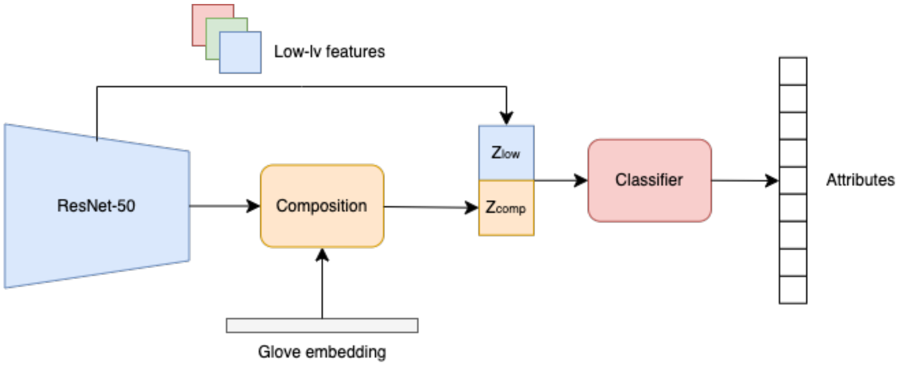


Figure 1: Attribute prediction model

follows:

$$f_{comp}(f_{img}(I), \phi_o) = f_{img}(I) \odot f_{gate}(\phi_o) \tag{1}$$

$$f_{gate}(\phi_o) = \sigma(W_{g2} \cdot ReLu(W_{g1}\phi_o + b_{g1}) + b_{g2}) \tag{2}$$

where $\odot$ is the channel-wise product, $\sigma(\cdot)$ is the sigmoid function, $f_{gate}(\phi_o) \in R^D$ is broadcast to match the feature map spatial dimension and is a 2-layer MLP.

5.3 Loss Function and Training Paradigm

Our final feature vector is the concatenation of the image-object composition and the low-level features from early blocks to improve accuracy for low-level attributes (color, material). The input to the classifier is $[Z_{low}, Z_{comp}]$, and we use a linear classifier with C output logit values followed by sigmoid.

5.3.1 Hard Labels Training

For the hard-labels training we apply the following re-weighted binary cross-entropy loss that takes data imbalance into account:

$$\mathcal{L}_{BCE}(Y, \hat{Y}) = \sum_{c=1}^{C} w_c(y_c \cdot p_c \cdot log(\hat{y}_c) + (1 - y_c) \cdot n_c \cdot log(1 - \hat{y}_c)) \tag{3}$$

where w_c, p_c, n_c are respectively the re-weighting factors for attribute c, its positive, and its negative example. We apply the same weighting method as Pham et al. [2021].

5.3.2 Class-Level Soft Labels Training

For the soft labels, we updated the $\mathcal{L}_{BCE}$ (Eq. 3) based on the human agreement. Every target y_c in the target vector Y was weighted by a factor k_c, which represents the scaled kappa agreement of the parent class of attribute c. The adopted scaling is a min-max normalization between 0.7 and 1 and it is chosen because the agreement score in Table 1 would have affected $\mathcal{L}_{BCE}$ too heavily. For example, a positive `size` attribute would be a target of 0.38 which would be incorrectly classified as negative.

5.3.3 Self-Knowledge Distillation Training

The architecture of the student network is identical to the architecture described in Section 5.2, with the exception of an additional loss function. We use the hard labels model as a teacher model, and include an

Table 2: Results

Methods	Full test set			Zero-shot subset		
	mAP	**mR@15**	**F1@15**	**mAP**	**mR@15**	**F1@15**
Hard labels	40.2	51.4	59.0	21.3	54.8	63.1
Soft labels	39.2	52.9	59.7	21.1	56.8	63.6
Distillation	40.4	52.4	59.5	21.3	55.4	62.4

additional distillation loss, $\mathcal{L}_{KD}$, to minimize output distribution distances between the teacher, T, and the student, S, via cross-entropy:

$$\mathcal{L}_{KD}(T, S) = -\sum_{c=1}^{C} \log s_c \cdot t_c. \tag{4}$$

The total loss becomes a linear combination of the binary cross-entropy and knowledge distillation losses, $\mathcal{L} = \mathcal{L}_{BCE} + \lambda\mathcal{L}_{KD}$, where we fix $\lambda = 1$.

5.4 Evaluation Metrics

We employ the following metrics to measure attribute prediction from different perspectives: mAP, $mR@15$ and $F1@15$ are computed for attribute classes, while SD_prob evaluates the model performance at attribute level, regardless of class type.

mAP: mean average precision over all classes. mAP was implemented as it is the primary metric used in Pham et al. [2021].

mR@15: mean recall over all classes at top 15 predictions in each image.

F1@15: harmonic mean of precision and recall at top 15 predictions in each image. $F1@15$ was used as $mR@15$ may be biased towards infrequent classes.

SD_prob: variance within attribute probabilities. SD_prob was used to access the ability of the models to predict negative and unlabelled attributes.

mAP, $mR@15$ and $F1@15$ are calculated on both the test set and the zero-shot subset defined in Section 4.1.

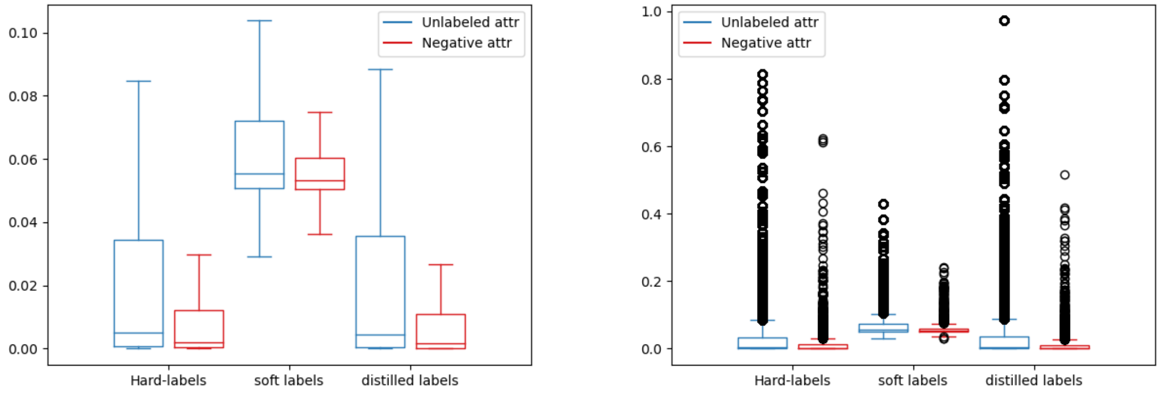
6 Results and Discussion

Table 2 shows the results on mAP, $mR@15$ and $F1@15$ for both the test set and the zero-shot subset. mAP is considered the primary evaluation metric, since it describes the quality of the model to rank correct images higher than the incorrect ones for each attribute label. Instead, $mR@15$ shows how well the model manages to output the ground truth positive attributes in its top 15 predictions in each instance. In addition, $F1@15$ is used to evaluate model performance taking into account both precision and recall.

Contrary to our hypothesis, we found no significant difference among the three models. In fact, the uncertainty introduced in the labels was not beneficial nor detrimental to the model performance. In addition, the knowledge distillation model also obtained comparable results to the other two, showing no difference between generating soft labels by exploiting human disagreement or using the predictions of another model.

A possible explanation for these results resides in the weightings added to the $\mathcal{L}_{BCE}$ (Eq. 3) to compensate for data imbalance. This weighting scheme already introduces modifications to the loss that might cover the effect of other additional weights, like the one introduced to generate soft labels, or of the additional loss for knowledge distillation.

As expected, the mAP values on the generalization are lower than the one on the full test set. Yet we find that the performance of $mR@15$ and $F1@15$ is higher on the zero-shot subset. This may be due to a difference



(a)

(b)

Figure 2: Mean and Standard Deviation of Attributes

in attribute class distribution, with attribute classes that are easier for the model to identify found in the zero-shot subset compared to the whole test set. However, we did not observe a noticeable difference in model performance between attribute class types.

Based on our hypothesis, we defined the SD_prob metric expecting that the introduction of soft labels would improve the model predictions not just for positive attributes but also unlabelled and negative ones. In fact, we expected:

- negative attributes to be expressed by values close to 0;
- unlabeled data to gain or lose probability depending on whether they are present or not in the instance.

As a result, we expected the variance of the labels predicted by the soft and distilled label models to be larger for unlabeled data than for negative labels. The findings of this analysis are depicted in Figure 2.

As shown from the distribution of data points in Figure 2a, the SD of unlabeled attributes is twice as big as the negative one for all models: `0.1 vs 0.05`, `0.04 vs 0.02` and `0.1 vs 0.04` respectively. However, for both unlabeled and negative labels, all models show on average low prediction probabilities. This could be the result of the small dimension of the dataset used for training and test steps. Because of the limited number of instances, the models failed to learn to discriminate between unlabelled and negative labels while trying to maximise the positive ones. Another explanation is that the averages are quite low and close to each other because just a minority of the 620 attributes can be actually utilized to define each instance.

Indeed, the presence of outliers (Figure 2b) reveals that the models can identify most unlabeled attributes as negative, although some are ascribed to the instance with a probability greater than 70%. Adding human uncertainty to the data, contrary to our expectation, appears to improve the model discrimination for negative attributes but not for unlabeled ones. Self-knowledge distillation, on the other hand, may be beneficial in detecting previously non-annotated positive attributes (although the results are not statistically significant).

More generally, our results are inconsistent with previous findings on using human disagreement to improve model performance on multi-class classification. This could be due to the partially labeled nature of the VAW dataset, or the selection of VAW subsets we have made. It could also be that results obtained on single-label classification tasks do not generalize to multi-label classification. Finally, the method we used to generate class-level soft labels could also be a reason, however, the fact that self-knowledge distillation also obtained comparable results seems to discard this possibility. On the other hand, if none of the mentioned causes affect the predictions, new investigations on the topic of learning from human disagreement would be necessary, especially in the context of attribute prediction.

7 Conclusion and Future Work

We investigated the problem of learning from human uncertainty for the task of attribute prediction on the VAW dataset. We presented a comparative evaluation of the use of soft and hard labels for this multi-label supervised classification task, specifically hard labels, class-level soft labeling and self-knowledge distillation.

Our hypothesis was that incorporating human disagreement would make our model more robust and better in terms of generalization than the two models that did not use this additional information (hard label training and self-knowledge distillation). However, we found no significant difference between the results obtained from the three models. Our results are inconsistent with previous findings on other NLP and CV tasks, and show that further evidence is needed on learning from disagreement, especially for attribute prediction.

Regarding future works, the model we used is able to perform attribute prediction only after several pre-processing steps. These are time-consuming, since each picture must be cropped before being input into the ResNet50. As described in Chen et al. [2023], it is feasible to overcome the slow pre-processing speed by training a faster end-to-end model for object recognition and attribute prediction where the picture is successively processed by a visual encoder, object region proposal, and attribute classification.

In addition, it has been consistently demonstrated that network designs frequently contain redundant neurons, which raises the likelihood of overfitting and optimisation issue [Hu et al., 2016]. As a result, future studies may focus on evaluating the model's architecture performing parametric or structural pruning. So that we may learn how different training procedures (hard labels, distilled labels, and soft labels) impact not only the model's performance but also the overall activity of its neurons.

References

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13018–13028, June 2021.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470, jan 2022. ISSN 1076-9757. doi:10.1613/jair.1.12752. URL `https://doi.org/10.1613/jair.1.12752`.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions, 2017. URL `https://openreview.net/forum?id=HkCjNI5ex`.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. *When Does Label Smoothing Help?* Curran Associates Inc., Red Hook, NY, USA, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi:10.1109/CVPR.2016.308.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL `https://arxiv.org/abs/1503.02531`.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/furlanello18a.html`.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi:10.3115/v1/P14-2083. URL `https://aclanthology.org/P14-2083`.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177, Oct. 2020. doi:10.1609/hcomp.v8i1.7478. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7478.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1176. URL https://aclanthology.org/N19-1176.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019. doi:10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1171. URL https://aclanthology.org/N18-1171.

Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 10, 2022.

Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, 2016.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973. ISSN 0010-0285. doi:https://doi.org/10.1016/0010-0285(73)90017-0. URL https://www.sciencedirect.com/science/article/pii/0010028573900170.

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. *ArXiv*, abs/1804.04340, 2018.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. in proceedings of the ieee conference on computer vision and pattern recognition. pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. Ovarnet: Towards open-vocabulary object attribute recognition. *ArXiv*, abs/2301.09506, 2023.

Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *EMNLP*, page 1532–1543, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis. Deep neural networks predict category typicality ratings for images. In *CogSci*, 2015.

Supplementary Materials

1 Dataset Statistics

1.1 Dataset statistics

Table 1 shows the dataset statistics: the number of images for each dataset, the number of object classes and the average number of positive and negative attribute per instance.

In the dataset we observe the typical long-tail distribution of object categories with some objects like `wall` and `floor` being represented in more than 2000 images, and other categories like `nurse` and `sea foam` having fewer than 10 instances. For the zero-shot detection investigation, we randomly picked 380 instances with an object frequency less than 10 from the 1214 object categories. Fig. 1 depicts the object categories shared by the various sets.

Because VAW is partially labeled, we only evaluated the models on the annotated data after removing instances having neither positive or negative attributes. As a result, the test set was downsampled by 16 occurrences.

Table 1: Dataset statistics

	# Images	# Objects	Pos attributes	Neg attributes
Training set	8800	991	1.518	2.027
Validation set	500	250	1.568	1.932
Test set	1279	545	1.523	1.695
Total Dataset	10579	1214	1.522	1.982

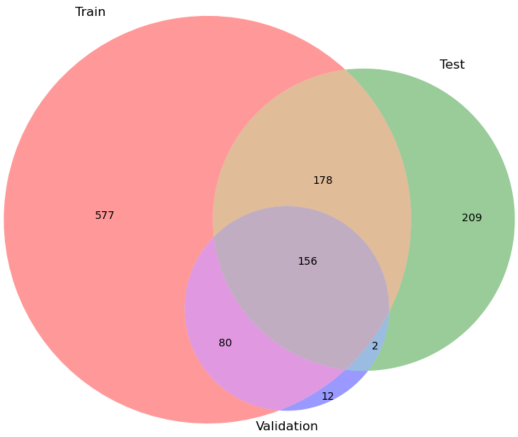
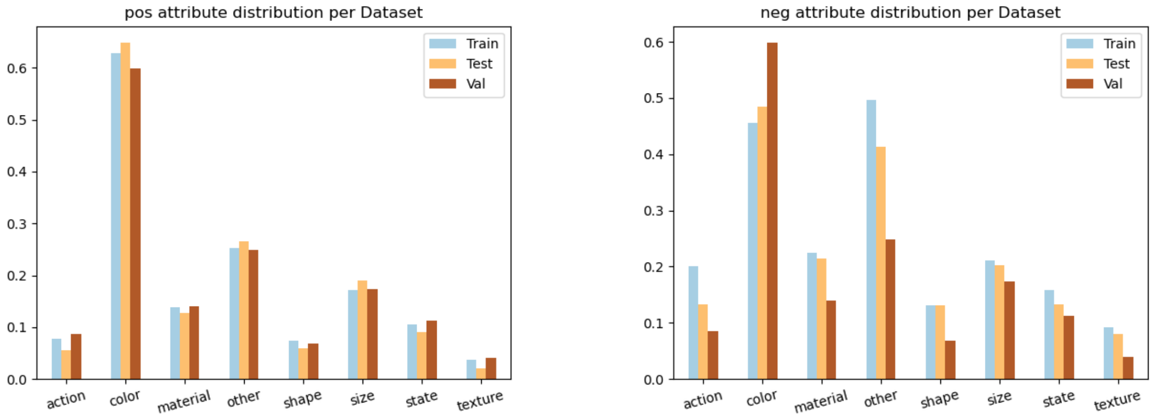


Figure 1: Venn diagram of object-categories distribution

1.2 Attributes statistics

Fig. 2a and 2b show the attribute class frequency per instance across different datasets. We can clearly observe a data imbalance problem. The frequency of the attributes classes vary a lot, rather than being roughly equal to each other. To prevent the model from overfitting the datapoints, this imbalance was addressed using weighted binary cross-entropy.



(a)

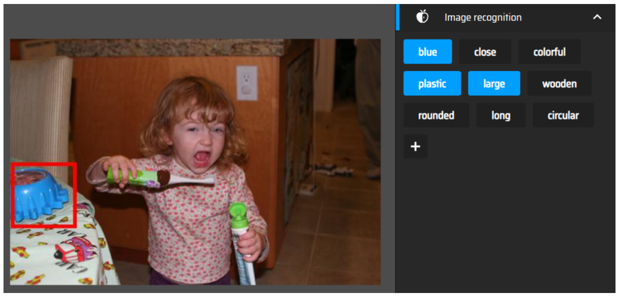
(b)

Figure 2: Attribute class frequency per instance

2 Soft-Labels Building Process

2.1 Soft-Labels from Human Disagreement

We collected human judgments for 100 random sampled instances from the VAW dataset. This sample contains at least one attribute for each parent category: `action`, `color`, `material`, `other`, `shape`, `size`, `state`, `texture`. We collected 5 human judgments per instance over the sample via online survey. The 5 annotators were asked to describe the object within a bounding box using the provided attributes as accurately as possible (with no time limit). For every attribute belonging to the list of positive attributes of an object, each annotator had to express a positive or a negative judgment about the appropriateness of the attribute w.r.t. the object (Figure 3 is an example). In the list of attributes, a random control was presented as an attention check, resulting in 279 annotated attributes, 116 of which without repetitions.
Agreement between the annotator's and the ground truth on control attributes was very low. This may indicate that there are positive attributes missed by the original annotators [Northcutt et al., 2021]. Agreement between annotator's themselves is also low. This might be caused by the fact that the control attributes were randomly chosen, changing the way in which annotators interpret them.



(a) online survey example

	Annotator #1	Annotator #2	Annotator #3	Annotator #4	Annotator #5
blue	yes	yes	yes	yes	yes
close *	no	no	yes	no	yes
plastic	yes	yes	yes	yes	yes
wooden (control)	no	no	no	no	no
large *	no	yes	yes	no	no
colorful *	no	yes	no	yes	yes
rounded	yes	yes	yes	yes	yes
long	no	no	no	no	no
circular *	yes	yes	yes	yes	no

(b) annotation example
$*$ = disagreement

Figure 3: Human annotation

Cohen's kappa agreement score among annotations for each class was used to create class-level soft labels. That is, instead of specifying how much each attribute resembles each category, we simply specify on which classes there is a higher disagreement using a class-level penalty. We propose to generate penalties by re-weighting the positive attributes based on the Cohen's kappa score k_c and assigning a value of 0.05 to the unlabeled data. The Cohen's kappa score k_c is obtained scaling kappa scores with a min-max normalization between 0.7 and 1.

As stated in Section 5.3.2, $\mathcal{L}_{BCE}$ (Eq. 3) was updated according to human disagreement. The parameter k_c was included to represent the scaled kappa agreement of the parent class of attribute c. The following equation shows $\mathcal{L}_{BCE}$ updated for soft labels:

$$\mathcal{L}_{BCE}(Y, \hat{Y}) = \sum_{c=1}^{C} w_c((y_c \cdot k_c) \cdot p_c \cdot log(\hat{y}_c) + (1 - (y_c \cdot k_c)) \cdot n_c \cdot log(1 - \hat{y}_c)) \tag{1}$$

2.2 Soft-Labels from Self-Knowledge Distillation

In addition to the cross-entropy loss, as described in Section 5.3.3, we also experimented with using Kullback-Leibler (KL) divergence as the distillation loss for a different measure of distance between the probability distributions of the teacher and student.

$$\mathcal{L}_{KL}(T, S) = -\sum_{c=1}^{C} t_c \cdot \log \frac{t_c}{s_c} \tag{2}$$

This did not materially affect the results, with a mAP of 40.35 compared to 40.38 for cross-entropy loss, so all other reported knowledge distillation results refer to knowledge distillation with cross-entropy loss.

3 Implementation Details

We used ResNet-50 He et al. [2016] pre-trained on ImageNet as the backbone, and use the output feature maps from ResNet blocks 2 and 3 as low-level features. For the object name embedding, we use the pre-trained GloVe Pennington et al. [2014] 100-d word embeddings. We do not fine-tune these word embeddings during training as we want our model to generalize to unseen objects during test time. The models are implemented in PyTorch Paszke et al. [2019] with the Adam optimizer Kingma and Ba [2014], weight decay of $1e^{-5}$, an initial learning rate of $1e^{-5}$ for the pre-trained ResNet and 0.007 for the rest of the model, and a batch size of 64. The models are trained on a single Nvidia RTX 3070. We follow the image augmentations of Pham et al. [2021], that is: random crop around the object bounding box, expanded by $\min(w, h) \cdot 0.3$; color jitter; random horizontal flip; and random grayscale on instances without color attributes. We train all models for 30 epochs, apply learning rate decay of 0.1 every time the mAP stops improving on the validation set for 2 epochs, and test on the model with best validation mAP.

4 Evaluation Metrics

4.1 mAP

mAP was introduced to accurately describe the quality of the model to rank correct images higher than incorrect ones for each attribute label. It was also implemented as it is the primary metric used in Pham et al. [2021].

mAP score is computed by taking the mean of the average precision of all C classes:

$$mAP = \frac{1}{C} \sum_{c} AP_c \tag{3}$$

$$AP_c = \frac{1}{P_c} \sum_{k=1}^{P_c} Precision(k, c) \cdot rel(k, c) \tag{4}$$

where P_c is the number of positive examples of class c, $Precision(k, c)$ is the precision of class c when retrieving the best k images, $rel(k, c)$ is the indicator function that returns 1 if class c is a ground-truth positive annotation of the image at rank k. As in Pham et al. [2021], mAP is only evaluated on annotated labels.

4.2 mR@15 and F1@15

$mR@15$ and $F1@15$ were introduced to measure the model performance considering the number of true positives for attribute class. We closely followed Pham et al. [2021] and Gong et al. [2013] to compute the precision, recall and F1 score. For each image, we consider the top 15 predictions of the model as its positive predictions. These predictions are then compared with the ground-truth annotations to compute the metrics.

4.3 SD_prob

SD_prob (variance within attribute prediction probabilities) was introduced to access the ability of the models to deal with negative and unlabelled attributes. SD_prob formula is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5}$$

where N is the number of positive, unlabelled and negative attributes for all the instances in the test set, x_i is the i^{th} model prediction, μ is the collective mean for a specific attribute tag (i.e. *positive, negative, unlabelled*).
We defined this metric with the hypothesis that unlabeled data predictions will have a larger standard deviation than negative data. This is because unlabeled data may gain or lose probability based on whether they are present or not in the instance.
Table 2 and Table 3 show some some label prediction statistics. As it can be seen, all models have very low prediction probabilities even for positive attributes. As already discussed within Section 6 of the paper, this might be due to an insufficient number of instances used for training and testing steps. Nevertheless, we carried out additional analyses to account for outliers or model misclassifications.

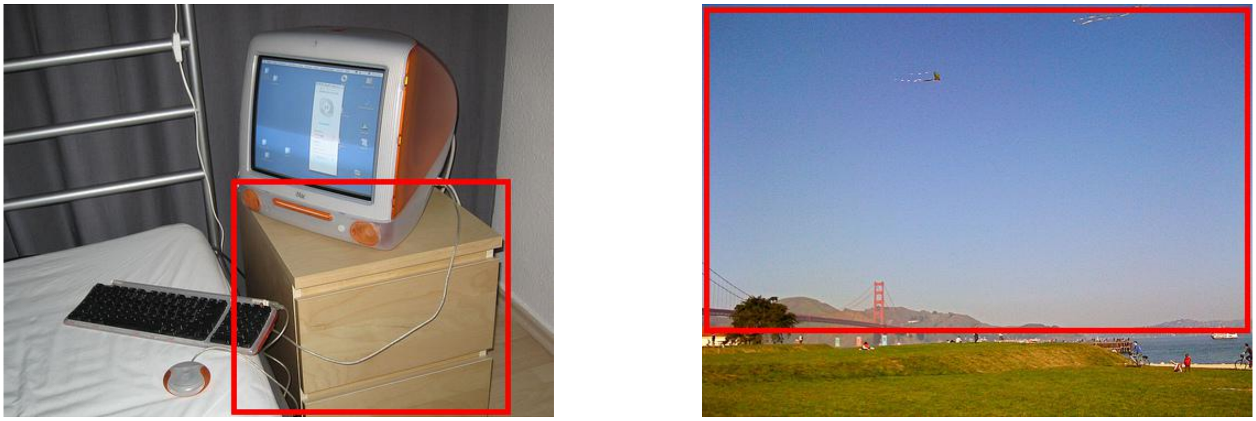
Table 2: Mean and standard deviation value within attribute prediction probabilities

	positive_attributes	unlabelled_attributes	negative_attributes
hard labels	0.0715 ± 0.1225	0.0436 ± 0.1008	0.0170 ± 0.0458
class-level soft labels	0.0847 ± 0.0496	0.0718 ± 0.0413	0.0598 ± 0.0202
distilled soft labels	0.0687 ± 0.1217	0.0454 ± 0.1070	0.0155 ± 0.0399

Table 3: Mean and standard deviation value of the gold-standard attribute labels

	positive_attributes	unlabelled_attributes	negative_attributes
hard labels	1.0 ± 0	0.0 ± 0.0	0.0 ± 0.0
class-level soft labels	0.7728 ± 0.0535	0.05 ± 0.0	0.0 ± 0.0

The model incorporating human uncertainty performs better in identifying negative attributes but is unable to associate some unlabelled attributes to instances as the other models do.
Figure 4 is an example of a negative and an unlabeled attribute's prediction probabilities. Different models evaluate the same attribute in different ways, and for these instances, attribute prediction via self-knowledge distillation seems to mimic human judgment more closely (though is not statistically significant).



(a)
neg: "white"
Image ID: 2326258
Instance ID: 2326258002
Object Name: nightstand
hard-pred: 0.10
soft-pred: 0.24
dist-pred: 0.06

(b)
unl: "blue"
Image ID: 2389203
Instance ID: 2389203004
Object Name: sky
hard-pred: 0.57
soft-pred: 0.48
dist-pred: 0.88

Figure 4: Attribute prediction example

5 VAW Predictions

Pham et al. [2021] shared with us the results obtained with their Strong Baseline + SCoNE (Supervised Contrastive Learning with Negative-label Expansion) model on the VAW test set.
We first considered using their model's prediction disagreement to strengthen the human one calculated over our annotators. However, as shown in Table 4, the Cohen's kappa agreement score among positive attributes predictions for each class is close to 0.5. As a result, we choose to rely solely on human disagreement and preserve a greater variance in the definition of the soft-labels.

Table 4: Kappa agreement score for each attribute class from Pham et al. [2021]

	action	color	material	other	shape	size	state	texture
Kappa score	0.51	0.57	0.58	0.52	0.54	0.52	0.53	0.52

Based on the assumption that disagreement may be related to the strength of a model's classification response to the category of interest [Lake et al., 2015], we then considered comparing their predictions to the results obtained with our models. However, since the instances in our test set were randomly selected, the overlap was limited to 170 examples. This, considering also that their model was trained with a far larger number of occurrences, forced us to discard this idea.