KNOWING WHO TO LISTEN TO: PRIORITIZING EXPERTS FROM A DIVERSE ENSEMBLE FOR ATTRIBUTE PERSONALIZATION

Shrenik Lad[1], *Bernardino Romera Paredes*[2], *Julien Valentin*[2], *Philip Torr*[2], *Devi Parikh*[1]

1. Virginia Tech 2. University of Oxford

ABSTRACT

Learning attribute models for applications like Zero-Shot Learning (ZSL) and image search is challenging because they require attribute classifiers to generalize to test data that may be very different from the training data. A typical scenario is when the notion of an attribute may differ from one user to another, e.g. one user may find a shoe *formal* whereas another user may not. In this case, the distribution of labels at test time is different from that at training time. We argue that due to the uncertainty in what the test distribution might be, committing to one attribute model during training is not advisable. We propose a novel framework for attribute learning which involves training an ensemble of diverse models for attributes and identifying experts from them at test time given a small amount of personalized annotations from a user. Our approach for attribute personalization is not specific to any classification model and we show results using Random Forest and SVM ensembles. We experiment with 2 datasets: SUN Attributes and Shoes and show significant improvements over baselines.

Index Terms— Attribute Learning, Ensemble Training, User Personalization

1. INTRODUCTION

Attributes are mid-level semantic properties of images like *shiny*, *furry*, *metallic*, etc. that are shared across categories. Recent research has shown that attributes have helped a variety of applications in computer vision ranging from object recognition [1, 2] and segmentation [3] to image description [4]. The fact that they are semantic and are shareable across categories makes them a suitable modality for novel forms of supervision, where humans can train visual models by conveying domain knowledge about the visual world. In image search, users can define a query using attributes (e.g., "find me *black*, *formal* shoes"), or provide feedback to a search engine to change the results along a certain attribute (e.g., "find me shoes that are more *shiny* than this") [5]. Zero-Shot Learning (ZSL) is another application where the task is to recognize previously unseen categories whose attribute descriptions are provided by a user but no training instances are available (e.g., recognizing a giraffe using just the knowledge that giraffes are *four-legged* and have *long necks*) [1].

Attribute classifiers are typically trained using images of certain training categories using annotations gathered from the crowd. These annotations are inexpensive and convenient to obtain via services like Amazon Mechanical Turk (AMT). Due to difference in perspective of various AMT workers, the annotations of multiple workers are averaged to get the final label for each training image. At test time, different users may have their personalized perception of an attribute, particularly for subjective attributes; a shoe that is *formal* for one user may not be *formal* for another user (different distribution of labels). As a result, generic attribute models trained from crowdsourced annotations may fail to satisfy individual user preferences at test time [6].

Since the test distribution is unknown at training time, we argue that committing to any one model for an attribute is not wise. Instead, learning an ensemble of multiple models for the attributes, where each model specializes in different parts of the feature space or learns different aspects of the attribute, is a better strategy to hedge against the uncertainty of the test distribution. We learn diverse models by training them on different subsets of the training (seen) *categories*. If absolutely no information is available about the test data, it is not possible to identify which models are likely to be a good fit for the particular task. In interactive applications like image search, a user can annotate a few images at test time to convey their notion of the attribute. Using this limited supervision, we find the experts from the ensemble that are well suited for that particular user and only use them for attribute prediction. We adopt a greedy technique for identifying experts where we pick the best performing models on the small set of labeled images from the user.

2. RELATED WORK

Ensemble methods like Random Forests (RF) and boosting learn multiple hypotheses for the same task and combine them for final prediction. Random forests have been shown to help various tasks in computer vision like image segmentation [7, 8], pose estimation [9] and edge detection [10]. Recently, CNN based ensembles have achieved state of the art performances in classification and detection challenges [11, 12, 13]. The CNNs in these ensembles are averaged for final predic-

tion. The different classifiers in an ensemble can be combined in a soft way via sum rule, product rule or majority vote [14, 15], or in a hard way by selecting only an individual classifier [16, 17] or a subset of classifiers [18]. Multiple Choice Learning [19] learns to predict multiple outputs, with the aim that the best output is close to ground truth. A re-ranker model whose parameters are also learned from training data, is used to pick one of the outputs. Our approach is also based on identifying a subset of models at test time but unlike previous approaches, our goal is to adapt the ensemble output to a new distribution.

Attributes have been used extensively for a variety of applications in computer vision such as object recognition [20, 21], scene understanding [22, 23], image description [4, 24], image search [5], clustering [25] and fine-grained recognition [26]. Many existing methods learn the attributes independently [23, 2, 1], often using linear models. To avoid learning coincidental correlations, [2] selects category-specific features that discriminate instances of the category containing an attribute from instances of the same category without that attribute. This approach does not apply to scenarios where attributes are defined at the category-level (e.g. all zebras are *striped*) as opposed to at the instance-level (e.g. this chair is *wooden*). With a similar goal of avoiding coincidental correlations, Jayaraman *et al.* [27] decorrelate attributes by encouraging feature competition between unrelated attribute groups and feature sharing within attributes of the same group. All of these existing works for attribute learning rely on a single attribute model to generalize to novel categories. Instead, we propose to learn multiple diverse models per attribute so that experts can be identified for individual users at test time.

Personalized applications like image search involve learning the user perception of relevance, and retrieving user specific results. Attributes being both machine detectable and human interpretable, serve as a useful mode of communication between humans and machines [5, 21, 25]. Attribute-based feedback (e.g., "find me shoes more *stylish* than this") has been shown to improve image search [5]. [25] proposes an approach to personalized constrained clustering where a user actively guides the machine by providing attribute-based explanations (e.g., "these two faces must be in the same cluster because both are *male* and *young*"). Studies have shown that humans tend to differ in their perception of subjective properties like *cool*, *formal*, etc. [28]. Kovashka and Grauman [6] introduce personalization in attribute learning by first training a generic attribute model using crowdsourced annotations and adapting the model to individual users at test time. Having a single model for an attribute limits the extent to which the model can adapt. Our approach instead identifies experts from a diverse pool of models at test time. We show that our approach outperforms [6].

Domain Adaptation is used to adapt a model trained from a source distribution to a different target distribution. We share similar goals. Unsupervised domain adaptation approaches [29, 30, 31] assume access to unlabeled data from the target distribution. Semi-supervised domain adaptation approaches [32, 33] use a few labeled samples from the test distribution (with or without the unlabeled pool), either during training [32] or at test time to re-train or update the model [34, 15]. The Adaptive SVM [34] technique used by Kovashka and Grauman for personalized attributes [6] is one such approach that adapts an SVM learned from a source distribution to a target distribution using a few labeled examples from the target distribution at test time. Note that our approach is not specific to a classifier. Hence, we can train an ensemble of diverse *adapted* models and then identify the (adapted) experts that are best suited for the test distribution (after all, some models will adapt better than others). In our experiments, we show that this approach outperforms a single adapted model as well as averaging the response of all adapted models in the ensemble.

3. APPROACH

We first describe our procedure to train diverse attribute models using Random Forests and SVMs, and then our approach for identifying experts from the ensemble given a small set of personalized attribute annotations at test time.

Let M be the number of binary visual attributes in our vocabulary that we want to learn. For training the attribute classifiers, we are given a set of training images X, where each image is annotated with attribute and category labels.

3.1. Training a Diverse Ensemble

We train multiple diverse models for each attribute in the vocabulary. These are *generic* models trained with a fixed set of training categories and crowdsourced annotations. Each model in the ensemble is trained with a different subset of the training data. Specifically, we train each model with a randomly sampled subset of the training *categories* (as opposed to training instances). This is done to encourage more diversity where each attribute model learns a notion of the attribute relevant to its own training categories. The categories are sampled with replacement.

Let p_{tm} denote the score/probability for the attribute a_m given by the t^{th} model in its ensemble. When there are no personalized annotations available, the final score for attribute a_m can be computed by simply averaging the scores of all models. If x denotes the image and T is the number of diverse models in the ensemble, the attribute prediction of the ensemble is given by

$$P_m(x) = \frac{1}{T}\sum_{t=1}^{T} p_{tm}(x) \tag{1}$$

We train Random Forests and SVM based ensembles as described in the following sections.

3.1.1. Random Forests

Consider a randomized decision forest with T trees. Let f be the low-level feature vector of x. The training of classification trees involves growing the trees recursively by learning (θ, τ) pairs at each node, where f_θ denotes the feature on which a node is split and τ denotes the threshold.

Instead of training the attributes independently, we train the attributes jointly using multi-output Random Forests, where each tree predicts all M attributes. This is because each tree can learn different correlations between attributes along with specializing on different categories.

The (θ, τ) pairs are learned so as to maximize the information gain at each node. Since we are learning attributes jointly, the entropy computation is over the distribution of M attributes. Ideally, we would like to estimate the exact entropy of the joint distribution which has 2^M possible states, but estimating this is not feasible with any reasonable quantities of data. The joint entropy can be approximated in many ways, for e.g., sum of individual attribute entropies at a node (ℓ_1 norm) that assumes independence of attributes [35], or maximum of individual attribute entropies (ℓ_∞ norm) that assumes complete dependence between attributes. We use the ℓ_2 norm of the individual entropies as a tradeoff between the two.

$$H(X) = \sqrt{\sum_{m=1}^{M} H_m(X)^2} \tag{2}$$

$H_m(X)$ denotes the entropy of attribute a_m at the node, which can be computed simply by counting the number of instances where the attribute is present/absent. At the leaves of a tree t, we store the posterior probabilities $p_{tm}(x)$ for all M attributes, where $p_{tm}(x)$ indicates the probability of attribute a_m being present in a training instance x that falls in that leaf node. This is indexed by the tree t because each instance can fall in only one leaf of a tree. The probabilities are computed independently for each attribute from the training data present at the leaves.

3.1.2. SVM

We train an ensemble of T SVMs for each attribute, resulting in $T \times M$ models in total. Unlike RFs, here the attributes are trained independently which is commonly done in existing works [2, 1, 27]. We use linear kernel and the C parameter for each SVM is selected after cross-validation on a held-out set from the training categories. If w_{tm} represents the weight vector of the t-th SVM for attribute a_m, the score on image x from the t-th SVM is given by

$$p_{tm}(x) = {w_{tm}}^\intercal f(x) \tag{3}$$

3.2. Identifying Experts - Personalized Attributes

Eq. 1 gives equal importance to every model in the ensemble. However, different models may have different prediction accuracies depending on the user. Hence, we identify expert models from the ensemble for each user and only use them for prediction.

Having a set of T diverse models for attribute a_m and a few personalized labeled images from a user, we evaluate each model in the ensemble individually on the labeled set. The models are then sorted based on their performances on the labeled images.

Specifically, if $X_U = (f_U, y_U)$ denotes the labeled data of a user U for attribute a_m, where f denotes feature descriptors and y denotes ground-truth attribute labels, the error of model t for attribute a_m is:

$$l_m(t) = \Delta(p_{tm}(X_U), y_U) \tag{4}$$

where Δ is a prescribed loss function. Note that each attribute need not have a different set of labeled images X_U since one image can be annotated with multiple (possibly all) attributes by the user.

When a domain adaptation technique is available for the generic classifier (e.g., Adaptive SVM for SVMs), we first use X_U to adapt each model in the ensemble. The adapted models are then evaluated on X_U to compute $l_m(t)$. Since X_U is being used for both purposes, to avoid overfitting, we adopt a N-fold cross-validation technique. In each fold, the models are adapted using the labeled images from $N-1$ folds, and the predictions of the adapted models on the remaining fold are stored. After N folds, each model is adapted using the entire labeled X_U. The loss of each model on X_U is estimated to be the loss of the stored predictions. In this way, we estimate $l_m(t)$ on entire X_U.

We sort the models based on $l_m(t)$ to find a ranking of the diverse models for individual users. After identifying the ranking of the models, we use only the top K models' predictions. If the experts for a_m are indexed by e_k, the attribute prediction for a_m is computed as:

$$P_m(x) = \frac{1}{K} \sum_{k=1}^{K} p_{e_k m}(x) \tag{5}$$

Note that the baseline approach in Eq. 1 uses all T models. Our approach in Eq. 5 uses only the *Top K* experts.

4. EXPERIMENTS AND RESULTS

The generic models for the attributes are trained using the approach described in Sec. 3.1. At test time, a small amount of images annotated with personalized attribute labels are available using which we find experts as described in Sec. 3.2.

We compare the following approaches: **Single Generic:** A single generic model trained using all available training data is applied directly on the new users at test time. This does not require any personalized labeled data. **Ensemble Generic:** The ensemble of generic models (Sec. 3.1) is applied directly on the new users by averaging all models' predictions. This also does not require any personalized labeled data. **Top K - Generic:** Our approach of finding experts from

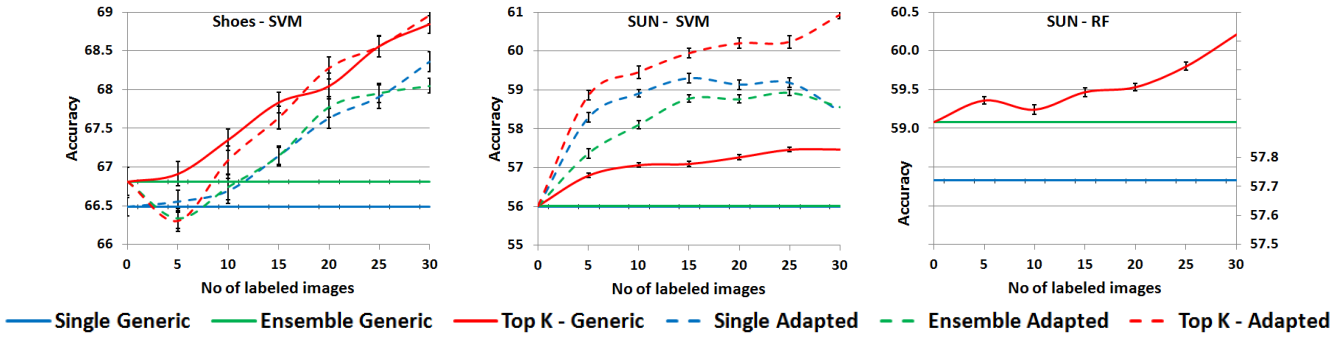

Fig. 1: User personalization results. Adaptive SVM based approaches are not applicable in case of RFs (right)

the generic ensemble and using top K models for attribute prediction. A small amount of personalized labeled data is used. **Single Adapted:** A single generic model is adapted for the new user using Adaptive SVMs [6, 34]. Personalized labeled data is used. **Ensemble Adapted:** Each generic model in the ensemble is adapted for the new user using Adaptive SVM. All adapted models are used by averaging the adapted models' predictions. Personalized labeled data is used. **Top K - Adapted:** Our approach of finding experts from the adapted models and using the top K models for attribute prediction. Personalized labeled data is used.

Note that the Adaptive SVM based approaches are applicable only in case of SVMs and not Random Forests. Comparing *Top K - Generic* with *Ensemble Generic* and *Top K - Adapted* with *Ensemble Adapted* shows the benefit of our expert selection approach. Comparing *Single Adapted* with *Single Generic* and *Ensemble Adapted* with *Ensemble Generic* shows the benefit of Adaptive SVM (not our contribution). Comparing *Ensemble Generic* with *Single Generic* and *Ensemble Adapted* with *Single Adapted* shows the benefit of using ensemble models over single models. Another natural baseline to consider is to learn attribute models from scratch using only the labeled data from new users. However, in our experiments we found that this performs significantly worse than a single generic model. This was also found by [6] for personalization of attributes.

We experiment with two datasets for the user personalization scenario: Shoes [36] and SUN Attributes [23].

Shoes dataset consists of images from 10 shoe categories like boots, flats, sneakers. The dataset contains 10 binary attributes like *formal*, *shiny*, *open*. The dataset contains around 14000 images in total with crowdsourced attribute annotations. [6] provides personalized attribute annotations on a small subset of 60 images for 10 different MTurk users. We reserve 30 images from these for adapting and identifying the experts and the remaining 30 images are used for evaluation. The generic attribute models are trained using 200 images from each shoe category. We use GIST and color histogram features available with the dataset.

SUN Attributes dataset consists of ∼14000 scene images belonging to more than 700 fine-grained categories. The dataset has 102 attributes like *natural*, *open*, *warm*, out of which 10 subjective attributes were chosen by [6] for personalized annotations. Similar to Shoes dataset, personalized annotations are available on a subset of 60 images for 5 different MTurk users. We follow the same protocol as in Shoes for evaluation. The generic attribute models are trained using 13000 images. We use GIST, HOG and SSIM features available with the dataset.

The generic models are adapted for each user separately at test time using their personalized annotations. We average attribute prediction accuracy across all users and attributes. The results are also averaged across 30 runs, where we sample different labeled images in each run. The metric used is normalized accuracy $= \frac{\text{TPR}+\text{TNR}}{2}$ (also used by [6]), where TPR and TNR denote true positive and true negative rates resp. We train 100 models in the ensemble for SUN Attributes and 50 models for Shoes (Shoes has relatively less training data).

Results: Fig.1 (left) shows results on Shoes dataset when SVMs are used. Our *Top K - Generic* approach performs better than both *Single Generic* and *Ensemble Generic* baselines, and our *Top K - Adapted* approach performs better than *Single Adapted* and *Ensemble Adapted*. This shows that we are able to identify user specific expert models from the diverse ensemble using limited user annotations. Results on SUN Attributes using SVMs can be found in Fig.1 (middle). The trends are similar to Shoes. Note that even though *Ensemble Adapted* performs little worse than *Single Adapted* in SUN-SVM, our Top K approach is able to identify experts and do better than the baselines.

Fig.1 (right) shows the results on SUN Attributes dataset when Random Forests are used. Again, Adaptive SVM based approaches are not applicable here. *Top K - Generic* does significantly better than *Single Generic* and *Ensemble Generic*. Nearly 2.5% gain is obtained over *Singe Generic* and 1% gain is obtained over *Ensemble Generic*. The results indicate that our approach is not specific to any classification model and gives significant gains over existing approaches that use a single generic model or a single adapted model.

Acknowledgements: This research is supported in part by an ARO YIP and an ARL award to D.P.

5. REFERENCES

[1] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling, "Learning to detect unseen object classes by betweenclass attribute transfer," in *In CVPR*, 2009. 1, 2, 3

[2] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, "Describing objects by their attributes," in *CVPR*, 2009. 1, 2, 3

[3] Shuai Zheng, Ming-Ming Cheng, Jonathan Warrell, Paul Sturgess, Vibhav Vineet, Carsten Rother, and Philip H. S. Torr, "Dense semantic image segmentation with objects and attributes," 2014, CVPR. 1

[4] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg, "Baby talk: Understanding and generating image descriptions," in *CVPR*, 2011. 1, 2

[5] Adriana Kovashka, "Whittlesearch: Image search with relative attribute feedback," 2012, CVPR. 1, 2

[6] Adriana Kovashka and Kristen Grauman, "Attribute adaptation for personalized image search," 2013, ICCV. 1, 2, 4

[7] Jamie Shotton, Matthew Johnson, and Roberto Cipolla, "Semantic texton forests for image categorization and segmentation," in *CVPR*, 2008. 1

[8] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests," 2008. 1

[9] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2013. 1

[10] Piotr Dollr and C. Lawrence Zitnick, "Structured forests for fast edge detection," 2013, ICCV. 1

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. 1

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *ArXiv e-prints*, 2014. 1

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*. 2012. 1

[14] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1998. 2

[15] Rajhans Samdani and Wen-tau Yih, "Domain adaptation with ensemble of feature groups," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2011. 2

[16] Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi, "Multi-output learning for camera relocalization," 2014, CVPR. 2

[17] Pyry K. Matikainen, Rahul Sukthankar, and Martial Hebert, "Model recommendation for action recognition," in *CVPR*, 2012. 2

[18] Marko Robnik-ikonja, "Improving random forests," in *ECML*. 2004. 2

[19] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli, "Multiple choice learning: Learning to produce multiple structured outputs," in *NIPS*, 2012. 2

[20] Jeff Donahue and Kristen Grauman, "Annotator rationales for visual recognition," 2011, ICCV. 2

[21] Amar Parkash and Devi Parikh, "Attributes for classifier feedback," 2012, ECCV. 2

[22] Ali Farhadi, Ian Endres, and Derek Hoiem, "Attribute-centric recognition for cross-category generalization.," in *CVPR*, 2010. 2

[23] Genevieve Patterson, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," 2012, CVPR. 2, 4

[24] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé, III, "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, EACL. 2

[25] Shrenik Lad and Devi Parikh, "Interactively guiding semi-supervised clustering via attribute-based explanations," in *ECCV*. 2014. 2

[26] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie, "Visual recognition with humans in the loop," 2010, ECCV. 2

[27] Dinesh Jayaraman, Fei Sha, and Kristen Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," 2014, CVPR. 2, 3

[28] William Curran, Travis Moore, Todd Kulesza, Weng-Keen Wong, Sinisa Todorovic, Simone Stumpf, Rachel White, and Margaret Burnett, "Towards recognizing "cool": Can end users help computer vision recognize subjective attributes of objects in images?," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, 2012, IUI. 2

[29] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Domain adaptation for object recognition: An unsupervised approach," 2011, ICCV. 2

[30] Minmin Chen, Kilian Q Weinberger, and John Blitzer, "Co-training for domain adaptation," in *NIPS*, 2011. 2

[31] John Blitzer, Ryan McDonald, and Fernando Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, EMNLP. 2

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," 2010, ECCV. 2

[33] Abhishek Kumar, Avishek Saha, and Hal Daume, "Co-regularization based semi-supervised domain adaptation," in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 478–486. 2

[34] Jun Yang, Rong Yan, and Alexander G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th International Conference on Multimedia*, 2007, MULTIMEDIA '07. 2, 4

[35] Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma, "Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages," in *Proceedings of the 22nd international conference on World Wide Web*, 2013. 3

[36] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih, "Automatic attribute discovery and characterization from noisy web data," 2010, ECCV. 4