Asking Friendly Strangers: Non-Semantic Attribute Transfer

Semantic visual attributes have allowed researchers to recognize unseen objects based on textual descriptions [1, 10, 13], learn object models expediently by providing information about multiple object classes with each attribute label [7, 14], interactively recognize fine-grained object categories [2, 17], and learn to retrieve images from precise human feedback [8, 9]. Recent ConvNet approaches have shown how to learn accurate attribute models through multi-task learning [4, 6] or by localizing attributes [15, 18]. However, deep learning with ConvNets requires a large amount of data to be available for the task of interest, or for a related task [12]. What should we do if we have a limited amount of data for the task of interest, and no data from semantically related categories? For example, let us imagine we have an entirely new domain of objects (e.g. deep sea animals) which is visually distinct from other objects we have previously encountered, and we have very sparse labeled data on that domain. Let us assume we have plentiful data from unrelated domains, e.g. materials, clothing, and natural scenes. Could we still use that unrelated data?

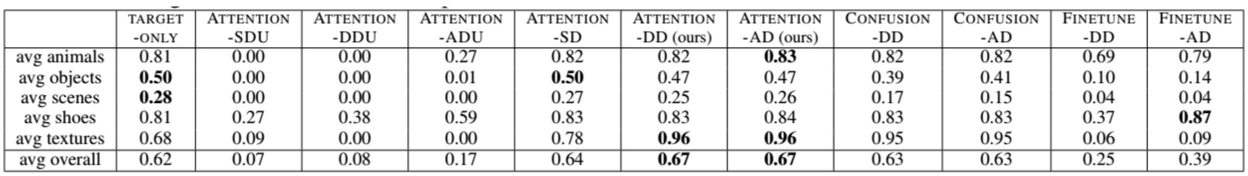
We examine how we can transfer knowledge from attribute classifiers on unrelated domains. For example, this might mean we want to learn a model for the animal attribute "hooved" from scene attribute "natural", texture attribute "woolen", etc. We define semantic transfer as learning a target attribute using the remaining attributes in that same data set as source models. This is the approach used in prior work [3, 5, 11]. In contrast, in non-semantic transfer (our proposed approach), we use source attributes from other datasets. We show that allowing transfer from diverse datasets allows us to learn more accurate models, but only when we intelligently select how to weigh the contribution of the source models. The intuition behind our approach is that the same visual patterns recur in different realms of the visual world, but language has evolved in such a way that they receive different names depending on which domain of objects they occur in.

We propose an attention-guided transfer network. Briefly, our approach works as follows. First, the network receives training images for attributes in both the source and target domains. Second, it separately learns models for the attributes in each domain and then measures how related each target domain classifier is to the classifiers in the source domains. Finally, it uses these measures of similarity (relatedness: W_{att}) to compute a weighted combination of the source classifiers, which then becomes the new classifier for the target attribute. We develop two methods, one where the target and source domains are disjoint, and another where there is some overlap between them. Importantly, we show that when the source attributes come from a diverse set of domains, the gain we obtain from this transfer of knowledge is greater than if only use attributes from the same domain.

We test our method on 272 attributes from five datasets of objects, animals, scenes, shoes, and textures, and compare it with several baselines in Table 1. Let D_i represent a domain and its attributes, and $D = U_{i=1}^{5} D_i$ be the union of all domains. We compare seven methods. The first are two ways of performing non-semantic transfer (our proposed approach) and the remaining ones are baselines.

- ATTENTION-DD, which is our multi-task attention network with D_i as our target domain and $D \backslash D_i$ as our source domains. We train five networks, one for each configuration of target/source.
- ATTENTION-AD, which is our multi-task attention network with D_i as our target domain and D as our source domains. We again train one network for each target domain. Some attributes on the source and target branches overlap, so we avoid transfer between an attribute and itself.
- ATTENTION-SD, which uses the same multi-task attention network but applies it on attributes from only a single domain D_i, for both the source and target branches. We again train five networks and avoid transfer between an attribute and itself.

- TARGET-ONLY, which uses data from the target attribute only, without any transfer from the source models.
- A replacement of the attention weights W_{att} with uniform weights. This results in baselines ATTENTION-SDU, ATTENTION-DDU and ATTENTION-ADU.
- [16] learns an invariant representation through a confusion loss and a domain classifier but no attention. This results in baselines CONFUSION-DD and CONFUSION-AD.
- Approaches FINETUNE-DD and FINETUNE-AD that fine-tune an AlexNet network using source data, then finetune those source networks again for the target domain [12].

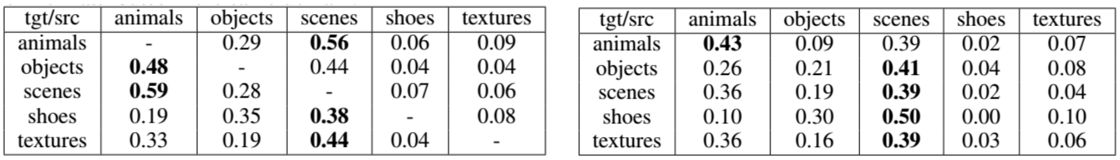


	TARGET -ONLY	ATTENTION -SDU	ATTENTION -DDU	ATTENTION -ADU	ATTENTION -SD	ATTENTION -DD (ours)	ATTENTION -AD (ours)	CONFUSION -DD	CONFUSION -AD	FINETUNE -DD	FINETUNE -AD
avg animals	0.81	0.00	0.00	0.27	0.82	0.82	**0.83**	0.82	0.82	0.69	0.79
avg objects	**0.50**	0.00	0.00	0.01	**0.50**	0.47	0.47	0.39	0.41	0.10	0.14
avg scenes	**0.28**	0.00	0.00	0.00	0.27	0.25	0.26	0.17	0.15	0.04	0.04
avg shoes	0.81	0.27	0.38	0.59	0.83	0.83	0.84	0.83	0.83	0.37	**0.87**
avg textures	0.68	0.09	0.00	0.00	0.78	**0.96**	**0.96**	0.95	0.95	0.06	0.09
avg overall	0.62	0.07	0.08	0.17	0.64	**0.67**	**0.67**	0.63	0.63	0.25	0.39

Table 1. Methods comparison using F-measure. Our approaches ATTENTION-DD and ATTENTION-AD outperform the other methods on average. Best results are **bolded** per row.

We also show qualitative results in the form of attention weights, which indicate what kind of information different target attributes borrowed. We extract and show the attention weights W_{att} for ATTENTION-DD (left) and ATTENTION-AD (right) in Table 2. Hence, for each target classifier i, we extract the weights W_{att_i} = (w_1; w_2; …; w_L) for the source classifiers. This procedure also verifies if ATTENTION-AD is primarily using transfer from attributes in the same domain, or attributes from disjoint domains with respect to the target. Due to a large number of attributes, we group attributes by their domain. Rows represent targets and columns sources.

Regarding ATTENTION-DD, the attention weights over the source classifiers are distributed among animals, objects, and scenes. We believe that shoes attributes are not very helpful for other domains because shoes images only contain one object. Further, textures are likely not very helpful because they are a low-level representation mainly defined by edges. Interestingly, we observe that the most relevant domain for animals, shoes, and textures is scenes, and scenes are *not closely related to any of these domains*. Similarly, the most meaningful domain for objects and scenes is animals, another *semantically unrelated source domain*. Regarding ATTENTION-AD, we observe that shoes and textures attributes do not benefit almost at all from other attributes in the same domain. On the other hand, objects, scenes, animals do benefit from semantically related attributes, but the overall within-domain model similarity is lower than 50%, again reaffirming our choice to allow non-semantic transfer.



tgt/src	animals	objects	scenes	shoes	textures
animals	-	0.29	**0.56**	0.06	0.09
objects	**0.48**	-	0.44	0.04	0.04
scenes	**0.59**	0.28	-	0.07	0.06
shoes	0.19	0.35	**0.38**	-	0.08
textures	0.33	0.19	**0.44**	0.04	-

tgt/src	animals	objects	scenes	shoes	textures
animals	**0.43**	0.09	0.39	0.02	0.07
objects	0.26	0.21	**0.41**	0.04	0.08
scenes	0.36	0.19	**0.39**	0.02	0.04
shoes	0.10	0.30	**0.50**	0.00	0.10
textures	0.36	0.16	**0.39**	0.03	0.06

Table 2. Attention weights summed per domain for our ATTENTION-DD (left) and ATTENTION-AD (right) approaches. Rows vs columns represent target vs source classifiers. The most relevant domains are **bolded** per row.

To conclude, while our target attributes come from well-defined and properly annotated datasets, our work demonstrates how non-semantic transfer can be used to learn attributes on novel domains where data is scarce, like the scenario discussed above. Our main contributions are an attention-guided transfer network and a study of transferability of attributes across semantic boundaries.

References

[1] Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[2] Branson, S.; Wah, C.; Schroff, F.; Babenko, B.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. In European Conference of Computer Vision (ECCV). Springer.

[3] Chen, C.-Y., and Grauman, K. 2014. Inferring analogous attributes. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[4] Fouhey, D. F.; Gupta, A.; and Zisserman, A. 2016. 3D shape attributes. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[5] Han, Y.; Yang, Y.; Ma, Z.; Shen, H.; Sebe, N.; and Zhou, X. 2014. Image attribute adaptation. IEEE Transactions on Multimedia.

[6] Huang, S.; Elhoseiny, M.; Elgammal, A.; and Yang, D. 2015. Learning hypergraph-regularized attribute predictors. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[7] Kovashka, A.; Vijayanarasimhan, S.; and Grauman, K. 2011. Actively selecting annotations among objects and attributes. In International Conference of Computer Vision (ICCV). IEEE.

[8] Kovashka, A.; Parikh, D.; and Grauman, K. 2015. Whittlesearch: Interactive image search with relative attribute feedback. International Journal of Computer Vision (IJCV).

[9] Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2011. Describable visual attributes for face verification and image search. Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

[10] Lampert, C.; Nickisch, H.; and Harmeling, S. 2009. Learning to Detect Unseen Object Classes By Between-Class Attribute Transfer. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[11] Liu, S., and Kovashka, A. 2016. Adapting attributes by selecting features similar across domains. In Applications of Computer Vision (WACV). IEEE.

[12] Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[13] Parikh, D., and Grauman, K. 2011. Relative attributes. In International Conference of Computer Vision (ICCV). IEEE.

[14] Parkash, A., and Parikh, D. 2012. Attributes for classifier feedback. In European Conference on Computer Vision (ECCV). Springer.

[15] Singh, K. K., and Lee, Y. J. 2016. End-to-end localization and ranking for relative attributes. In European Conference on Computer Vision (ECCV). Springer.

[16] Tzeng, E.; Hoffman, J.; Darrell, T.; and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. In International Conference on Computer Vision (ICCV). IEEE

[17] Wah, C., and Belongie, S. 2013. Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop. In Computer Vision and Pattern Recognition (CVPR). IEEE.

[18] Xiao, F., and Jae Lee, Y. 2015. Discovering the spatial extent of relative attributes. In International Conference on Computer Vision (ICCV). IEEE.