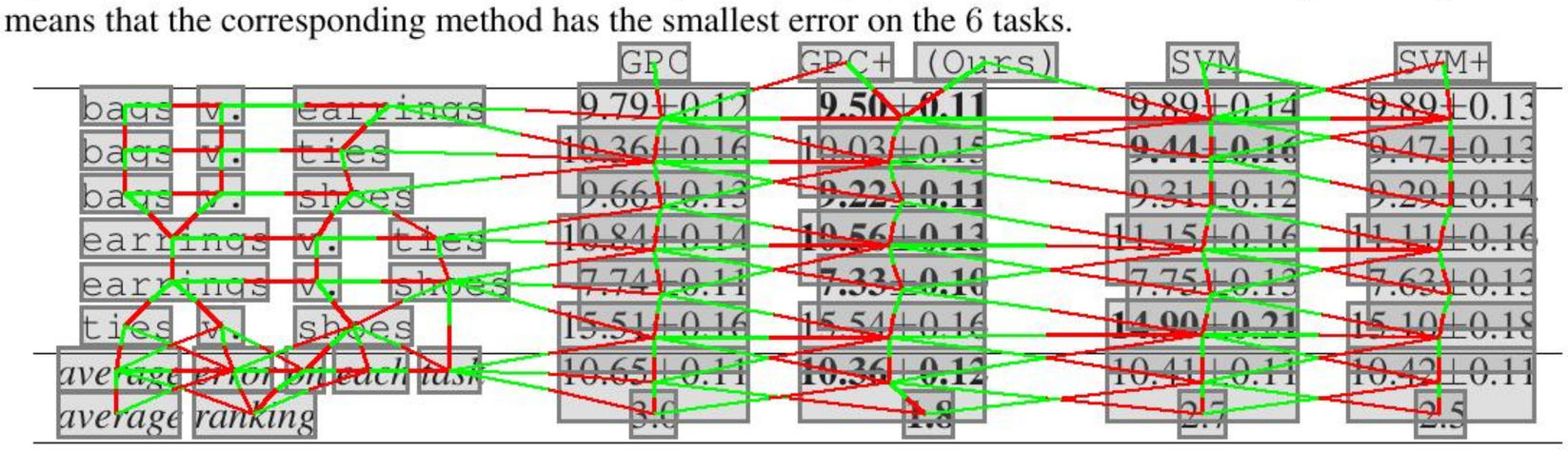
Table 1: Average error rate in % (the lower the better) on the Attribute Discovery dataset over 100 repetitions. We used **images** as the **original** domain and neural networks word-vector representation on **texts** as the **privileged** domain. The best method for each binary task is highlighted in **boldface**. An average rank equal to one means that the corresponding method has the smallest error on the 6 tasks.



[6], we generated 45 binary classification tasks for each pair of the 10 classes with 200 samples for training, 200 samples for validation, and the rest of samples for testing the predictive performance.

Neural networks on images as privileged information: Deep learning methods have gained an increased attention within the machine learning and computer vision community over the recent years. This is due to their capability in extracting informative features and delivering strong predictive performance in many classification tasks. As such, we are interested to explore the use of deep learning based features as privileged information so that their predictive power can be used even if we do not have access to them at prediction time. We used the standard *SURF* features [21] with 2000 visual words as the *original* domain and the recently proposed *DeCAF* features [25] extracted from the activation of a deep convolutional network trained in a fully supervised fashion as the *privileged* domain. The DeCAF features have 4096 dimensions. All features are provided with the AwA dataset[2]. We again performed PCA for dimensionality reduction in the original and privileged domains and only kept the top 50 principal components, as well as standardised the data.

Attributes as privileged information: Following the experimental setting of [6], we also used *images* as the *original* domain and *attributes* as the *privileged* domain. Images were represented by 2000 visual words based on SURF descriptors and attributes were in the form of 85 dimensional predicted attributes based on probabilistic binary classifiers [24]. As previously, we also performed PCA and kept the top 50 principal components in the original domain and standardised the data.

The results of these experiments are shown in Figure 2 in terms of pairwise comparisons over 45 binary tasks between GPC+ and the main baselines, GPC and SVM+. The complete results with the error of each method GPC, GPC+, SVM, and SVM+ on each problem are relegated to the supplementary material. In contrast to the results on the attribute discovery dataset, on the AwA dataset it is clear that GPC outperforms SVM in almost all of the 45 binary classification tasks (see the supplementary material). The average error of GPC over 4500 (45 tasks and 100 repeats per task) experiments is much lower than SVM. On the AwA dataset, SVM+ can take advantage of privileged information – be it deep belief DeCAF features or semantic attributes – and shows significant performance improvement over SVM. However, GPC+ still shows the best overall results and further improves the already strong performance of GPC. As illustrated in Figure 1 **(right)**, the privileged information modulates the slope of the probit likelihood function differently for easy and difficult examples: easy examples gain slope and hence importance whereas difficult ones lose importance in the classification. In this dataset we analysed our experimental results using the multiple dataset statistical comparison method described in [26].[3] The results of the statistical tests are summarised in Figure 3. When DeCAF attributes are used as privileged information, there is statistical evidence supporting that GPC+ *performs best* among the four methods, while when the semantic attributes are used as privileged information, GPC+ still performs best but there is not enough evidence to reject that GPC+ performs comparable to GPC.

[2] http://attributes.kyb.tuebingen.mpg.de

[3] Note that we are not able to use this method on the results of the attribute discovery dataset in Table 1 because the number of methods compared (*i.e.*, 4) is almost equal to the number of tasks or datasets (*i.e.*, 6).