Appendix

A. Orthogonal signatures

In this section we prove that if each attribute signature of the training classes is orthogonal to each attribute signature of the test classes, that is, if for each $i \in \{1 \dots z\}$, $j \in \{1 \dots z'\}$, $\langle s_i, s'_j \rangle = 0$, then the right hand side term in eq. (10) becomes bigger than one.

To make the explanation clearer let us denote by $x \in \mathbb{R}^d$ any training instance in the original feature space, and similarly let $x' \in \mathbb{R}^d$ be any test instance. Then, by applying eq. (8) using the training signature s_i, and test signature s_j we have

$$\widetilde{x}_i = \text{vec}\left(x s_i^\top\right) \in \mathbb{R}^{da}$$

$$\widetilde{x}'_j = \text{vec}\left(x' s_j'^\top\right) \in \mathbb{R}^{da}$$

Note that because of the orthogonality assumption between training and test signatures the following holds true:

$$\left\langle \widetilde{x}_i, \widetilde{x}'_j \right\rangle = \text{trace}\left(x s_i^\top s'_j x'^\top\right) = 0. \tag{11}$$

Eq. (11) implies that in the new feature space any training instance is orthogonal to any test instance. Because of that, the following lemma becomes useful.

Lemma 1. *Let us consider $\mathcal{H}$ be the hypothesis space composed of all linear classifiers. Then given two orthogonal sets $\mathcal{P}$, $\mathcal{Q}$, there exists a hypothesis $f \in \mathcal{H}\Delta\mathcal{H}$ which separates them.*

Proof. Let us consider any couple of points $p \in \mathcal{P}$, $q \in \mathcal{Q}$ with the only condition that they are not zero. We define

$h(x) = \text{sign}\left((p+q)^\top x\right)$, and

$h'(x) = \text{sign}\left((p-q)^\top x\right)$.

For any point $p' \in \mathcal{P}$,

$$\begin{aligned} h(p') &= \text{sign}\left((p+q)^\top p'\right) \\ &= \text{sign}\left(p^\top p'\right) \\ &= \text{sign}\left((p-q)^\top p'\right) \\ &= h'(p') \end{aligned}$$

given that by definition p' and q are orthogonal. Similarly, for any point $q' \in \mathcal{Q}$,

$$\begin{aligned} h(q') &= \text{sign}\left((p+q)^\top q'\right) \\ &= \text{sign}\left(q^\top q'\right) \\ &= -\text{sign}\left((p-q)^\top q'\right) \\ &= -h'(q') \end{aligned}$$

Therefore $f \in \mathcal{H}\Delta\mathcal{H}$ associated to functions $h, h' \in \mathcal{H}$ will be positive if and only if the input is in $\mathcal{Q}$. □

As a consequence of Lemma 1, when the orthogonality assumption holds, the right hand side term in eq. (10) becomes bigger than 1, so that the bound is vacuous as one would expect. One illustrative instance of this case happens when $S = [B, 0_{a,c},]$, and $S' = [0_{a,b}, C]$ for some non-zero matrices $B \in \mathbb{R}^{a \times b}$, $C \in \mathbb{R}^{a \times c}$. In that case, the set of attributes that describe the training classes are completely different from the ones describing the test classes, thus no transfer can be done.

B. More synthetic experiments

In this section we add further synthetic experiments that complement the ones reported in Sec. 5.1, and give support to some claims made thorough the paper.

Firstly, we focus on the question discussed in Sec. 4.2 by empirically assessing the performance of our approach as a function of the similarity between S and S'. We use the trace norm of the product, $\left\|S^\top S'\right\|_{\text{Tr}}$, as a way to measure the similarity between S and S'. We have considered here a similar experimental setting as the one described in Sec. 5.1, and we have fixed the number of attributes to be 5. We have repeated the experiment 10000 times, and the results are reported in Fig. 4. We have also calculated the correlation between the multiclass accuracy and $\left\|S^\top S'\right\|_{\text{Tr}}$, which is 0.1877. In the figure we observe that when $\left\|S^\top S'\right\|_{\text{Tr}}$ is low, there is not any instance of high multiclass accuracy. When $\left\|S^\top S'\right\|_{\text{Tr}}$ is high, any result is possible, from low to high multiclass accuracy. One way to interpret this is that high correlation between the signatures in S and S' is necessary to obtain a high performance, but that is not the only factor for success. Other factors may be related, for example, the performance of the approach may be inversely related to the similarity of the test signatures between each other (and correspondingly the similarity of the training signatures), because it is difficult to distinguish between classes when the attributes that describe them are not discriminative.

Secondly, we extend the last experiment of Sec. 5.1 in which we studied the extreme case where some attributes provide no information at all about the classes at hand. Let us recall that in that experiment we randomly selected ψ attributes, and tweaked them. In the present experiment we assess whether our model diminishes the importance of these attributes through the learned weights V.

Let us define by $\mathcal{A}$ the set of all attributes, with cardinality a. From this set $\mathcal{A}$ we randomly sample ψ misleading attributes, creating the set $\Psi \subseteq \mathcal{A}$. These attributes will be artificially tweaked in order to test the robustness of our approach. Similarly we denote by $\Phi = \mathcal{A} \setminus \Psi$ the remainder set of $a - \psi$ informative attributes. We denote by $V_\Psi \in \mathbb{R}^{d \times \psi}$ and $V_\Phi \in \mathbb{R}^{d \times (a-\psi)}$ the matrix consist-

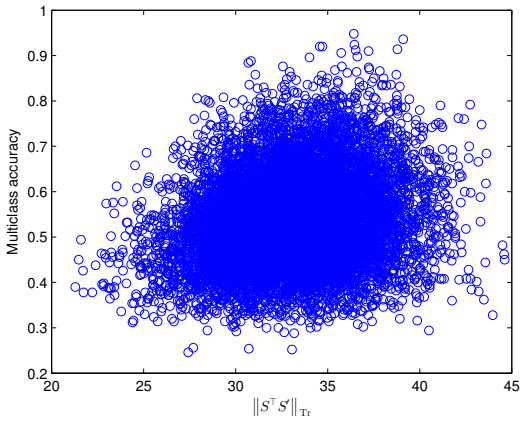

Figure 4. Performance of ESZSL with respect to the similarity between S and S'.

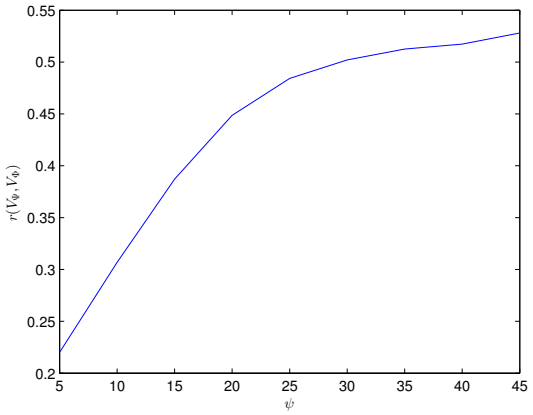

Figure 5. Values of $r(V_\Psi, V_\Phi)$ for V obtained using ESZSL, when varying the number of corrupted attributes ψ.

ing on the columns of V associated to the attributes in Ψ and Φ respectively. If our model is learning to filter out misleading attributes, it should learn a V such that V_Ψ has a significantly lower magnitude than V_Φ. That is exactly what we quantify in this experiment. We consider the relative importance of the weights in Ψ with respect to the Φ as $r(V_\Psi, V_\Phi) = \frac{(a-\psi)\|V_\Psi\|^2_{\text{Fro}}}{\psi\|V_\Phi\|^2_{\text{Fro}}}$, where $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm. Note that $r(V_\Psi, V_\Phi) = 1$ if the average squared ℓ_2 norm of the columns in Ψ is similar to the average squared ℓ_2 norm of the columns in Φ, that is, if the model is not capable of distinguishing between misleading and meaningful attributes.

We have used a similar experimental setting as the one in Sec. 5.1, that is, we fixed $a = 100$, and varied ψ in the range $[5, 10, \ldots, 45]$. We repeated 200 trials for each value of ψ and report the average results in Fig. 5. We can see that the values obtained are much lower than 1, implying that our model is able to discriminate meaningful attributes. Furthermore, we see that $r(V_\Psi, V_\Phi)$ grows as the number of misleading attributes increases, although it seems to plateau for high values of ψ.

C. More real experiments

In this section we present some additional experiments on real datasets.

C.1. Attributes prediction

The focus of our model is on maximising the multiclass accuracy among the classes at hand. However, as a byproduct of the learning process, we can also use V as a way to predict attributes. In this experiment we check whether these attribute predictors are effective, or on the contrary, the gain in zero-shot performance comes at the expense of attribute prediction. In order to do so, we compare the described option with a simple approach that learns an SVM for each attribute directly. The results are reported in table 4.

Mean Average Precision	AwA	aPY	SUN
Learning attributes directly	**56.95%**	**30.78%**	**79.36%**
Using $X^\top V$ from ESZSL	50.73%	29.51%	68.53%

Table 4. Comparison between SVM (Learning attributes directly), and ESZSL, for attributes prediction, using mean average precision as a measure.

The gain in ZSL performance comes at the expense of attribute prediction. This may be because our approach tends to neglect the attributes that are unreliable or useless for class prediction, whereas in attribute prediction all are considered equally important. These results are in the same vein as the ones reported in (Akata et al., 2013).

C.2. Sample results using SUN dataset

In this section we present some qualitative results in order to gain insights about when ESZSL succeeds and when it fails. In Fig. 6 we show some samples of the SUN test set organised in a confusion matrix, according to the predictions made by ESZSL. White cells within this matrix indicate that there is no error made between the corresponding two classes.

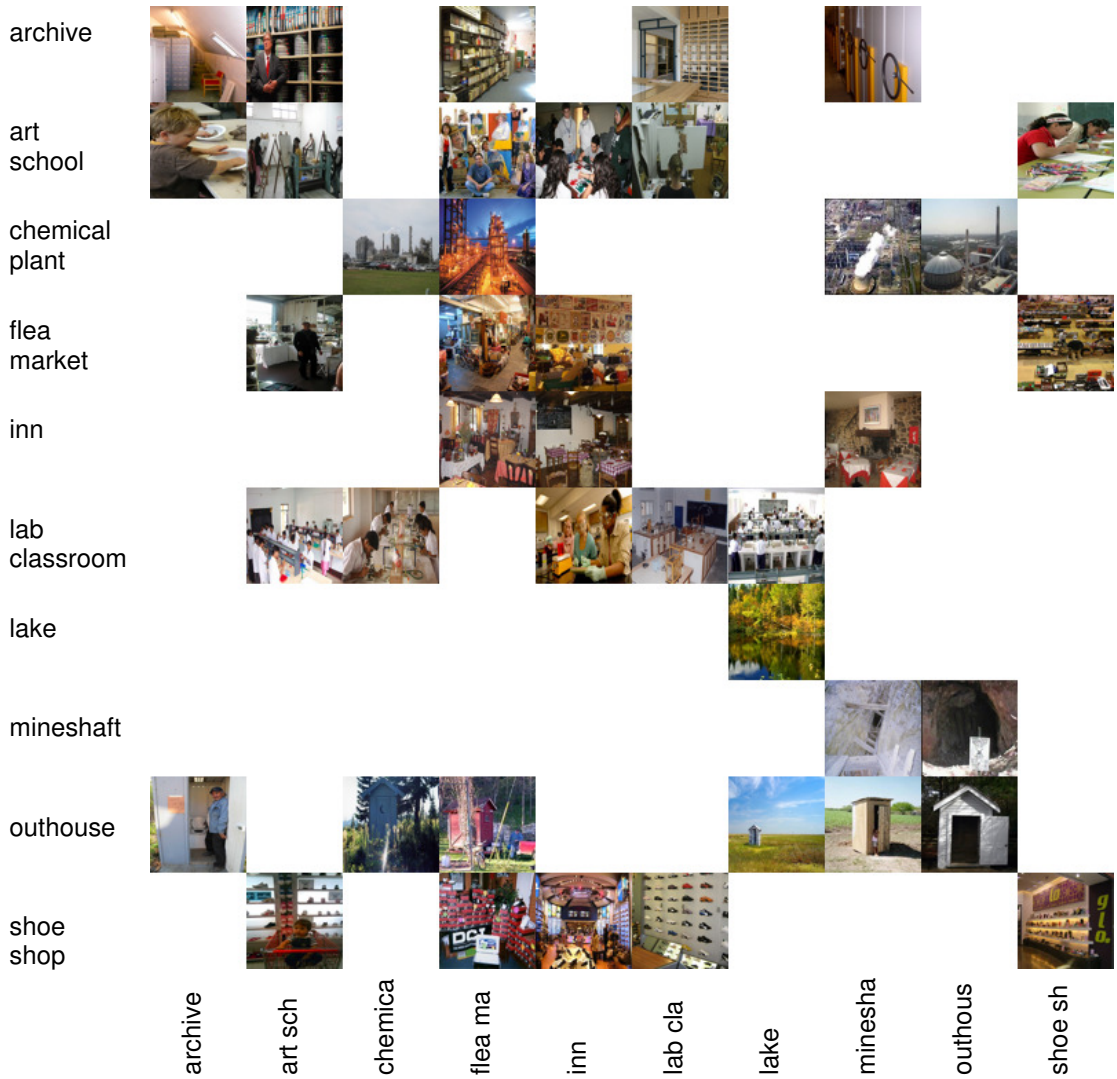

Figure 6. Samples of the SUN test set in a confusion matrix. Rows represent ground truth classes, columns represent predictions.