Supplementary Material for the Paper: Probabilistic AND-OR Attribute Grouping for Zero-Shot Learning

A IMPLEMENTATION AND TRAINING DETAILS

The weights W were initialized with orthogonal initialization (Saxe et al., 2014). The loss in Eq. (6) was optimized with Adam optimizer (Kingma & Ba, 2015). We used cross-validation to tune early stopping and hyper-parameters. When the learning rate is too high, the number of epochs for early-stopping varies largely with the weight seed. Therefore, we chose a learning rate that shows convergence within at least 40 epochs. Learning rate was searched in [3e-6, 1e-5, 3e-5, 1e-4, 3e-4]. From the top performing hyper-parameters, we chose the best one based on an average of additional 3 different seeds. Number-of-epochs for early stopping, was based on their average learning curve. For β, λ, L2 regularization params, we searched in [0, 1e-8, .., 1e-3].

For learning soft groups, we also tuned the learning rate of V in [0.01, 0.1, 1], of ζ in [1, 3, 10], and when applicable, the number of groups K in [1, 10, 20, 30, 40, 60], or semantic prior ψ in [1e-5, .., 1e-2]. We tuned these hyper-params by first taking a coarse random search, and then further searching around the best performing values.

To comply with the mutual-exclusion approximation (2), if the group sum $\sum_{m \in G_k} p(a_m = T|z)$ is larger than 1, we normalize it to 1. We do not normalize if the sum is smaller than 1 in order to allow LAGO to account for the complementary case. We apply this normalization only for the LAGO-Semantic variants, where a prior knowledge about grouping is given.

After selecting hyper-parameters with cross-validation, models were retrained on both the training and the validation classes.

A.1 EVALUTATION METRIC

We follow Xian et al. (2017) and use a class-balanced accuracy metric which averages correct predictions independently per-class before calculating the mean value:

$$acc_{\mathcal{Z}} = \frac{1}{|\mathcal{Z}|} \sum_{z=1}^{|\mathcal{Z}|} \frac{\text{\# of correct predictions in } z}{\text{\# of samples in } z}. \quad \text{(A.1)}$$

B LEARNED SOFT-GROUP ASSIGNMENTS (Γ)

We analyzed the structure of learned soft group assignments ($\Gamma_{m,k} = p(m \in G_k)$) for LAGO-K-Soft, initialized by a uniform prior. We found two types of interesting structures:

First, we find that the learned Γ tends to be sparse: with 2.5% non-zero values on SUN, 8.7% on AWA2 and 3.3% on CUB. As a result, the learned model has small groups, each with only a few attributes. Specifically, Γ maps each attribute to only a single group on SUN (K=40 groups) and CUB (K=30), and to 2-3 groups on AWA2 (K=30 groups).

Second, we tested which attributes tend to be grouped together, and found that the model tends to group anti-correlated attributes. To do this, we first quantified for each pair of attributes, how often they tend to appear together in the data. Specifically, we estimated the occurrence pearson-correlation for each pair of attributes across samples (CUB, SUN) or classes (AWA2). Second, we computed the grouping similarity of two attributes as the inner product of their corresponding rows in Gamma, and considered an attribute pair to be grouped together if this product was positive (note that rows are very sparse). Using these two measures, we observed that the model tends to group anti-correlated attributes. This is consistent with human-based grouping, whose attribute are also often anti correlated (red foot, blue foot). In SUN, 45% of attribute-pairs that are grouped together were anti-correlated, compared to 23% of the full set of pairs. (AWA2 38% vs 5% baseline, CUB 16% vs 10% baseline). These differences were also highly significant

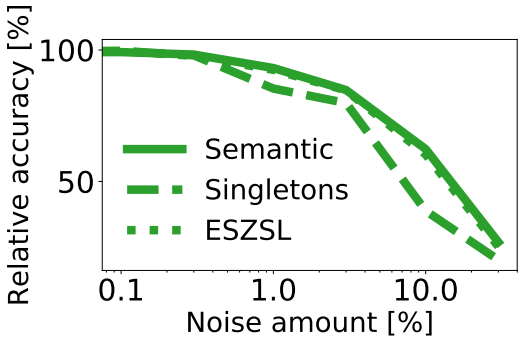

Figure A.1: Robustness to salt & pepper noise. The relative accuracy on CUB of three models as a function of ration of injected noise to class-level description $p(a_m|z)$. Values are averages over 5 noise-seeds. LAGO-Semantic-Hard and ESZSL show a similar sensitivity to noise, while LAGO-Singletons is more sensitive due to its all-AND structure. The relative accuracy is calculated against each model own zero-noise baseline.

statistically (Kolmogorov-Smirnov test p-value¡3e-3)

C ROBUSTNESS TO NOISE

We tested LAGO-Semantic-Hard, LAGO-Singletons and ESZSL with various amount of salt & pepper noise (Figure A.1) injected to class-level description $p(a_m|z)$ of CUB. While LAGO-Semantic-Hard and ESZSL show a similar sensitivity to noise, LAGO-Singletons is more sensitive due to its all-AND structure.

D DETAILED DERIVATION

D.1 $p(a_m|g_{k,z}{=}T)$ EQUALS $p(a_m|Z{=}z)$

Here we explain why Eq. (A.2) below is true.

$$p(a_m|g_{k,z}{=}T) = p(a_m|Z{=}z), \tag{A.2}$$

It is based on the definition of $g_{k,z}$: $g_{k,z}$ is the classifier of z based on $\mathbf{a}_k$. Therefore $p(g_{k,z}|\mathbf{a}_k){=}p(z|\mathbf{a}_k)$, and by marginalization we get: (*) $p(g_{k,z}=T, a_m){=}p(z, a_m)$, (**) $p(g_{k,z}=T){=}p(Z=z)$. Next, using conditional probability chain rule on (*), yields

$$p(a_m|g_{k,z}{=}T)p(g_{k,z}{=}T) = p(a_m|Z=z)p(Z=z) \quad . \tag{A.3}$$

Then, (**) transforms (A.3) to the required equality:

$$p(a_m|g_{k,z}{=}T) = p(a_m|Z=z) \quad . \tag{A.4}$$

Intuitively, the right side of (A.4), is the probability of observing a_m for a class z, like $p(stripes|zebra)$. This is the same probability of observing the attribute given the class while focusing on its respective group, namely $p(a_m = T|g_{k,z} = T) = p(stripes|$*focus on zebra pattern*$)$.

D.2 DERIVATION OF GROUP CONJUNCTION:

This derivation is same as in DAP (Lampert 2009), except we apply it at the group level rather than the attribute level. We denote $g_{1,z} \dots g_{K,z}$ by $\mathbf{g}_z$ and approximate the following combinatorially large sum:

$$p(Z{=}z|\mathbf{x}){=}\sum_{\mathbf{g}_z\in\{T,F\}^K} p(Z{=}z|\mathbf{g}_z)p(\mathbf{g}_z|\mathbf{x}) \quad . \tag{A.5}$$

First, using Bayes (A.5) becomes

$$\sum_{\mathbf{g}_z\in\{T,F\}^K} \frac{p(\mathbf{g}_z|Z{=}z)p(Z{=}z)}{p(\mathbf{g}_z)} p(\mathbf{g}_z|\mathbf{x}) \tag{A.6}$$

Second, we approximate $p(\mathbf{g}_z|Z{=}z)$ to be

$$p(\mathbf{g}_z|Z{=}z) = \begin{cases} 1, & \text{if } g_{1,z}{=}T \dots g_{K,z}{=}T \\ 0, & \text{otherwise} \end{cases} \tag{A.7}$$

which transforms (A.6) to

$$p(Z{=}z|\mathbf{x}) \approx p(Z{=}z)\frac{p(g_{1,z}{=}T \dots g_{K,z}{=}T|\mathbf{x})}{p(g_{1,z}{=}T \dots g_{K,z}{=}T)} \tag{A.8}$$

Third, we approximate the numerator of (A.8) with the assumption of conditional independence of groups given an image (by observing an image we can judge each group independently),

$$p(g_{1,z}{=}T \dots g_{K,z}{=}T|\mathbf{x}) \approx \prod_{k=1}^{K} p(g_{k,z}{=}T|\mathbf{x}) \tag{A.9}$$

Fourth, we approximate the denominator of (A.8) to its factored form $p(g_{1,z} = T \dots g_{K,z} = T) \approx \prod_{k=1}^{K} p(g_{k,z}{=}T)$, and with (A.9) we arrive at:

$$p(Z{=}z|\mathbf{x}) \approx p(Z{=}z) \prod_{k=1}^{K} \frac{p(g_{k,z}{=}T|\mathbf{x})}{p(g_{k,z}{=}T)} \quad . \tag{A.10}$$

D.3 A DERIVATION OF SOFT GROUP MODEL

Here we adapt LAGO to account for soft group-assignments for attributes, by extending the within-group part of the model. We start with partitioning $p(g_{k,z} = T|\mathbf{x})$ to a union (OR) of its contributions, repeated below for convenience,

$$p(g_{k,z}|\mathbf{x}) = p(g_{k,z}, \bigcup_{m\in G_k} a_m = T|\mathbf{x}) + p(g_{k,z}, \tilde{a_k} = T|\mathbf{x}), \tag{A.11}$$

	CUB	AWA2	SUN
DAP	40.0	46.2	39.9
ALE	54.9	62.5	**58.1**
ESZSL	53.9	58.6	54.5
SYNC	55.6	46.6	56.3
SJE	53.9	61.9	53.7
DEVISE	52.0	59.7	56.5
ZHANG2018	48.7 - 57.1	58.3-70.5	57.8-61.7
LAGO-SINGLETONS	54.5	63.7	57.3
LAGO-K-SOFT	55.3	59.7	57.5
LAGO-SEMANTIC-HARD	58.3	60.4	47.1
LAGO-SEMANTIC-SOFT	57.8	64.8	48.0
LAGO (CROSS-VALIDATION)	**57.8**	**64.8**	57.5

Table A.1: Test accuracy for all the variants of LAGO on three benchmark datasets, averaged over 5 random initializations of model weights. Standard-error-of-the-mean (S.E.M) is $\sim 0.1\%$ for the hard groups variants and $\sim 0.4\%$ for the soft-groups variants.

$p(a_m)$	$P(\tilde{a_k}\|\mathbf{x})$	ATTRIBUTES SUPERVISION	CUB	AWA2
UNIFORM	CONST	IMPLICIT	**52.85**	**60.53**
UNIFORM	CONST	EXPLICIT	52.17	60.17
UNIFORM	DEMORGAN	EXPLICIT	51.75	57.93
UNIFORM	DEMORGAN	IMPLICIT	48.41	49.54
PER-ATTRIBUTE	DEMORGAN	EXPLICIT	47.71	53.32
PER-ATTRIBUTE	CONST	EXPLICIT	42.68	52.05
PER-ATTRIBUTE	CONST	IMPLICIT	39.31	51.88
PER-ATTRIBUTE	DEMORGAN	IMPLICIT	35.3	37.21

Table A.2: Ablation experiments: Validation accuracy (in %) for CUB and AWA2, for combinations of model-design variants, with the semantic hard-grouping of LAGO. Results are given in descending order based on CUB. *Uniform* vs *Per-attribute* relates to taking a uniform prior for $p(a_m)$. *Const DeMorgan* relates to setting a constant value for approximating the complementary attribute $p(\tilde{a_k}|\mathbf{x})$ vs an approximation derived by De-Morgan's rule. *Implicit* vs *Explicit* relates to setting a zero weight ($\alpha = 0$) for the loss term of the attribute supervision. Namely, attributes are learned implicitly, since only class-level super vision is given. The uniform prior on $p(a_m)$ has the largest impact, second is the usage of a constant value for $p(\tilde{a_k}|\mathbf{x})$, and the last relates to nulling the attribute supervision loss. See details on Section 4.4

and instead, treat the attribute-to-group assignment $(m \in G_k)$, as a probabilistic assignment, yielding:

$$p(g_{k,z}|\mathbf{x}) = p(g_{k,z}, \bigcup_{m=1}^{|\mathcal{A}|}(m \in G_k, a_m = T|\mathbf{x})) + p(g_{k,z}, \tilde{a_k} = T|\mathbf{x}), \tag{A.12}$$

Note that the attribute-to-group assignment $(m \in G_k)$ is independent of the current given image $\mathbf{x}$, class z or the True / False occurrence of an attribute a_m. Repeating the mutual exclusion approximation (2) yields,

$$p(g_{k,z}|\mathbf{x}) \approx \sum_{m=1}^{|\mathcal{A}|} p(g_{k,z}, m \in G_k, a_m = T|\mathbf{x}). \tag{A.13}$$

Using the independence of $(m \in G_k)$, yields

$$p(g_{k,z}|\mathbf{x}) \approx \sum_{m=1}^{|\mathcal{A}|} p(m \in G_k) p(g_{k,z}, a_m = T|\mathbf{x}). \tag{A.14}$$

Defining $\Gamma_{m,k} = p(m \in G_k)$, yields:

$$p(g_{k,z}|\mathbf{x}) \approx \sum_{m=1}^{|\mathcal{A}|} \Gamma_{m,k} p(g_{k,z}, a_m = T|\mathbf{x}). \tag{A.15}$$

As in section 3, using the Markov chain property $\mathcal{X} \to \mathcal{A} \to \mathcal{G}$ and $p(g_{k,z} = T|a_m) = \frac{p(a_m|z)p(g_{k,z}=T)}{p(a_m)}$ results with Eq. (5), repeated below:

$$p(g_{k,z} = T|\mathbf{x}) \approx p(g_{k,z} = T) \sum_{m=1}^{|\mathcal{A}|} \Gamma_{m,k} \frac{p(a_m = T|z)}{p(a_m = T)} p(a_m = T|\mathbf{x}) \tag{A.16}$$

D.3.1 APPROXIMATING THE COMPLEMENTARY TERM

With soft groups, the complementary term is defined as

$$\tilde{a_k} = \Big(\bigcup_{m=1}^{|\mathcal{A}|} (m \in G_k, a_m = T|\mathbf{x})) \Big)^c \qquad \text{(A.17)}$$

To approximate $p(\tilde{a_k} = T|z)$ we can use De-Morgan's rule over a factored joint conditional probability of group-attributes. I.e.

$$p(\tilde{a_k} = T|z) \approx \prod_{m=1}^{|\mathcal{A}|} (1 - p(m \in G_k, a_m = T|z)) = \prod_{m=1}^{|\mathcal{A}|} (1 - \Gamma_{m,k} p(a_m = T|z)), \quad \text{(A.18)}$$

where the latter term is derived by the independence of $(m \in G_k)$

D.4 DAP, ESZSL AS SPECIAL CASES OF LAGO

Two extreme cases of LAGO are of special interest: having each attribute in its own singleton group ($K = |A|$), and having one big group over all attributes ($K = 1$).

Consider first assigning each single attribute a_m to its own singleton group ($K = |A|$ and $m = k$). We remind that we defined $G'_k = G_k \bigcup \tilde{a_k}$. Therefore, G'_k has only two attributes $\{a_m, \tilde{a_k}\}$, which turns the sum in Eq. (4), to a sum over those elements:

$$p(z|\mathbf{x}) = p(z) \prod_{k=1}^{K} \Big[\frac{p(a_m{=}T|z)}{p(a_m{=}T)} p(a_m{=}T|\mathbf{x}) + \frac{p(\tilde{a_k}{=}T|z)}{p(\tilde{a_k}{=}T)} p(\tilde{a_k}{=}T|\mathbf{x}) \Big]. \quad \text{(A.19)}$$

In a singleton group, the complementary attribute $\tilde{a_k}$ becomes $\tilde{a_k} = a_m^c$, and therefore
$\tilde{a_k}{=}T \Leftrightarrow a_m{=}F$. This transforms (A.19) to:

$$p(z|\mathbf{x}) = p(z) \prod_{m=1}^{|A|} \Big[\frac{p(a_m{=}T|z)}{p(a_m{=}T)} p(a_m{=}T|\mathbf{x}) + \frac{p(a_m{=}F|z)}{p(a_m{=}F)} p(a_m{=}F|\mathbf{x}) \Big] \quad . \quad \text{(A.20)}$$

This formulation is closely related to DAP (Lampert et al., 2009), where the expert annotation $p(a_m{=}T|z)$ is thresholded to $\{0, 1\}$ using the mean of the matrix U as a threshold, and denoted by a_m^z. Applying a similar threshold to Eq. (A.20) yields

$$p(z|\mathbf{x}) = p(z) \prod_{m=1}^{|A|} \Big[\frac{a_m^z}{p(a_m{=}T)} p(a_m{=}T|\mathbf{x}) + \frac{(1 - a_m^z)}{p(a_m{=}F)} p(a_m{=}F|\mathbf{x}) \Big] \quad \text{(A.21)}$$

Reducing Eq. (A.21), by taking only the cases where it is non-zero for its two parts, gives the posterior of DAP

$$p(z|\mathbf{x}) = p(z) \prod_{m=1}^{|A|} \frac{p(a_m{=}a_m^z|\mathbf{x})}{p(a_m{=}a_m^z)} \quad . \quad \text{(A.22)}$$

This derivation reveals that in the extreme case of $K = |A|$ singleton groups, LAGO becomes equivalent to a soft relaxation of DAP.

At the second extreme, consider the case where all attributes are assigned to a single group, $K = 1$. Taking a uniform prior for $p(z)$ and $p(a_m)$, and writing $p(a_m = T|\mathbf{x})$ using the network model $\sigma(\mathbf{x}^\top W)$, transforms Eq. (4) to:

$$p(z|\mathbf{x}) \propto \sum_{m=1}^{|A|} \sigma(\mathbf{x}^\top W) p(a_m = T|z). \qquad \text{(A.23)}$$

This can be viewed as a 2-layer architecture: First map image features to a representation in the attribute dimension, then map it to class scores by an inner product with the supervised entries of attributes-to-classes $U_{m,z} = p(a_m = T|z)$. This formulation resembles ESZSL, which uses a closely related 2-layer architecture: $Score(z|\mathbf{x}) = \mathbf{x}^\top WU$, where W first maps image features to a representation in the same attribute dimension, and then map it to class scores, with an inner product by the same attributes-to-classes entries $U_{m,z} = p(a_m = T|z)$. LAGO differs from ESZSL in two main ways: (1) The attribute-layer in LAGO uses a sigmoid-activation, while ESZSL uses a linear activation. (2) LAGO uses a cross-entropy loss, while ESZSL uses mean-squared-error. This allows ESZSL to have a closed-form solution where reaching the optimum is guaranteed.

This derivation reveals that at the extreme case of $K = 1$, LAGO can be viewed as a non-linear variant that is closely related to ESZSL.

References

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*. IEEE, 2009.

A.M. Saxe, J. L McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *ICLR*, 2014.

Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, 2017.