A. Additional negative results

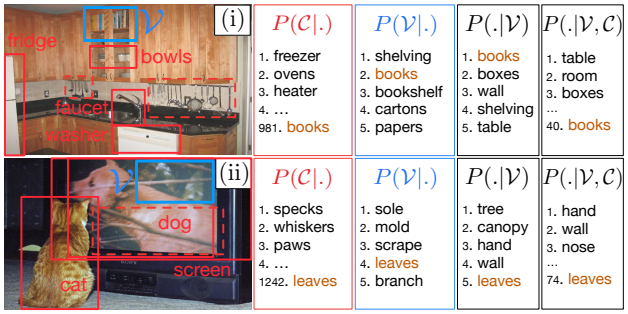

Figure 5. Qualitative analysis: negative examples where the use of the context leads to degraded predictions, i.e. examples where model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ is worse than the simpler model $\mathrm{M}(\mathcal{V})$ ($\mathcal{T}$ classes only).

As explained in Section 5.3, using contextual information can sometimes degrade predictions. We provide here additional examples, when an object occurs in an environment in which it is unexpected. For example, Figure 5 shows a picture of a kitchen where the object of interest to be predicted is "books". Given only the surrounding environment, predicted objects are logically related to the environment of a kitchen ("freezer", "oven", ...), and the correct label is badly ranked (because it is unexpected in such an environment). However, the model $\mathrm{M}(\mathcal{V})$ retrieves the correct label, given only the region of interest. Finally, integrating contextual information in the final model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ leads to worse performances over $\mathrm{M}(\mathcal{V})$.

B. Generalized ZSL

In the previous sections, retrieval in done only among classes of the domain of interest, this is the classical zero-shot learning setting. We now report results obtained when both source and target object classes exist in the retrieval space: this setting amounts to *generalized zero-shot learning*. Results are reported in Table 3.

Table 3. Evaluation of various information sources, with varying levels of supervision. Generalized ZSL setting. MFR scores in %. δ_C is the relative improvement (in %) of $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ over $\mathrm{M}(\mathcal{V})$.

		Target domain $\mathcal{T}$			Source domain $\mathcal{S}$		
	p_{sup}	10%	50%	90%	10%	50%	90%
	Domain size	4358	2421	484	484	2421	4358
	Random	*100*	*100*	*100*	*100*	*100*	*100*
Models	$\mathrm{M}(\varnothing)$	39.6	26.3	16.9	6.6	8.68	10.9
	$\mathrm{M}(\mathcal{V})$	21.0	11.8	6.9	0.9	2.3	3.5
	$\mathrm{M}(\mathcal{C}_{S_H})$	28.6	15.0	10.7	3.5	3.9	4.4
	$\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$	18.2	9.4	6.0	0.8	1.8	2.4
	$\delta_{\mathcal{C}}$	*13.4*	*20.2*	*13.4*	*13.8*	*24.4*	*31.5*

C. MRR and top-k performances

ZSL models are usually evaluated with recall@k or MRR (mearn reciprocal rank, i.e. harmonic mean). However, the metrics are not optimal to evaluate our models for two reasons:

- Theoretically, recent research points out that RR is not an interval scale and thus MRR should not be used (Fuhr, Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum 2017 ; Ferrante et al. Are IR evaluation measures on an interval scale? ICTIR 2017).
- Practically, we make the size of the target domain vary (10%, 50%, 90%). MRR and top-k scores cannot be compared across these scenarios (e.g. top-5 among 100 entities is not comparable to top-5 among 1000)

Therefore, as explained in Section 4.2 we used MFR (mean first relevant): the arithmetic mean of rank numbers (linearly rescaled to have 100% for random model and 0% for perfect model). FR is an interval scale and thus can be averaged.

However, we report here top$k$ and MRR scores in Table 4.

Table 4. Recall@k ($k \in \{1, 5, 10\}$) (in percentage) and MRR scores (in percentage). $p_{\text{sup}} = 50\%$.

	Target domain $\mathcal{T}$				Source domain $\mathcal{S}$			
	Recall @			MRR	Recall @			MRR
	1	5	10		1	5	10	
Random	<.1	0.2	0.4	<.1	<.1	0.2	0.4	<.1
$\mathrm{M}(\varnothing)$	3.2	11.7	16.3	7.8	5.7	17.9	24.9	12.5
$\mathrm{M}(\mathcal{V})$	14.7	33.5	43.2	24.0	36.3	63.8	73.1	48.8
$\mathrm{M}(\mathcal{C}_{S_H})$	5.9	17.8	25.4	11.9	17.3	43.7	56.7	29.9
$\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$	15.0	34.7	44.7	24.7	41.6	70.6	78.6	54.2