Relational Proxies: Emergent Relationships as Fine-Grained Discriminators

Supplementary Material

Abhra Chaudhuri[1] **Massimiliano Mancini**[2] **Zeynep Akata**[2,3,4] **Anjan Dutta**[5]*
[1] University of Exeter [2] University of Tübingen [3] MPI for Informatics
[4] MPI for Intelligent Systems [5] University of Surrey

1 Additional Experiments

1.1 Fine-grained performance boost on ImageNet subsets over SotA

We compare our method to TransFG [2], the SotA FGVC method on Dogs ImageNet. We summarize our findings in Table 1, which shows that our method provides state-of-the-art performance boost in the fine-grained setting over vanilla relation-agnostic encoders. Δ_1 and Δ_2 denote the perfomance boost achieved by an FGVC method over relation-agnostic encoders in the coarse-grained and fine-grained settings respectively.

Relational features play a much more significant role in distinguishing fine-grained categories than coarse-grained ones. This is because most coarse-grained classes can be distinguished by local/global features alone, and would not require relational information. However, for fine-grained classes, the cross-view relationships often happen to be the only available discriminator. Thus, a learner not leveraging the same would suffer from the information gap (Section 3.2 in the main manuscript), not providing any significant boost over a relation-agnostic encoder. Our method, by capturing the cross-view relationships, is able to bridge this information gap.

Method	Tiny ImageNet	Δ_1	Dogs ImageNet	Δ_2	$\Delta_2 - \Delta_1$
Relation-Agnostic Encoder	88.75		91.30		
TransFG [2]	88.85	0.10	92.30	1.00	0.90
Relational Proxy (Ours)	88.91	0.16	92.75	1.45	**1.29**

Table 1: Comparison of coarse *vs.* fine-grained accuracy gains over a relation-agnostic encoder.

1.2 Permutation invariance of AST

For our method to be robust to changes in pose and relative orientation of local object parts, we require the Attribute Summarization Transformer (AST) to be permutation invariant. We achieve the same by eliminating position embeddings [4] from our AST. We test the validity of our requirement by comparing the classification accuracy of Relational Proxies having ASTs with and without position embeddings [4]. We summarize our findings in Table 2, which shows that making the AST permutation invariant in fact plays a role in enhancing the performance of our model.

Given the low inter-class variation of the cultivar datasets, parts of leaves from different classes could appear the same under changes in orientation, making a permutation sensitive model mistake it for a different class. For this reason, the AST without position embeddings (permutation invariant)

*A. Chaudhuri is with the Department of Computer Science at the University of Exeter. M. Mancini and Z. Akata are with the Cluster of Excellence Machine Learning at the University of Tübingen. A. Dutta is with the Institute for People-Centred AI at the University of Surrey.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

performs significantly better (compared to other benchmarks) than the one with position embeddings (permutation sensitive).

Method	Benchmark				Cultivar	
	FGVC Aircraft	**Stanford Cars**	**CUB**	**NA Birds**	**Cotton**	**Soy**
w/ Position Embeddings	95.11	96.15	91.82	91.09	68.77	50.15
w/o Position Embeddings	**95.25**	**96.30**	**92.00**	**91.20**	**69.81**	**51.20**

Table 2: Effect of position embeddings on the permutation invariance of the Attribute Summarization Transformer (AST).

1.3 Evaluation with VGG-16 Backbone

To ensure that our method has no backbone specific dependency, we perform evaluations with VGG-16 [5] backbone and report our findings in Table 3. As the numbers show, our method remains stable across backbones, significantly outperforming SotA methods that report performances with VGG-16 backbones as well.

Method	FGVC Aircraft	CUB
MaxEnt [1]	78.08	77.02
MMAL [6]	87.00	83.75
Ours (Relational Proxies)	**91.20** ±0.03	**88.13** ±0.01

Table 3: Comparison of our method with state-of-the-art using VGG-16 backbone.

2 Qualitative Results

2.1 Importance of Relational Information

Figure 1 shows examples of classes that cannot be separated by global or local information alone. The cross-view relational information serves as the strongest discriminator for such classes. For example, Black-footed Albatross, Laysan Albatross and the Sooty Albatross (denoted in red, dark blue and orange respectively), share a large number of local attributes and have similar overall appearances, but have differing geometries. For this reason, as can be observed from the low-dimensional visualization of their embeddings obtained via UMAP [3], they are only separable based on their relational features, and not by their global or local features. Additionally, Figure 2 shows that such classes becomes separable as the model learns to incorporate the relational information as part of the learning process.

2.2 Relation-Agnosticity of Relational Proxies

Figure 3 shows UMAP visualizations of global and local embeddings for instances of a single class, obtained from a fully trained Relational Proxy model. It provides empirical evidence for our theoretical result in Lemma 3, *i.e.*, f will produce relation-agnostic representations if the downstream objective is cross-entropic in nature. As can be seen, despite using the same set of proxies for the global and the local views, they get mapped to disjoint locations in the representation space. The distance between the clusters of global and local views is proportional to the information gap (Proposition 1), which is separately being learned by the relational encoder ξ (Proposition 2). However, some global embeddings can still be seen to overlap with the cluster of the locals. This happens with images for which the information provided by the global view becomes redundant after collectively knowing the set of local views. The global view does not provide any additional information and thus can be merged with the local views with no information loss (while maintaining the requirement of k-distinguishability).

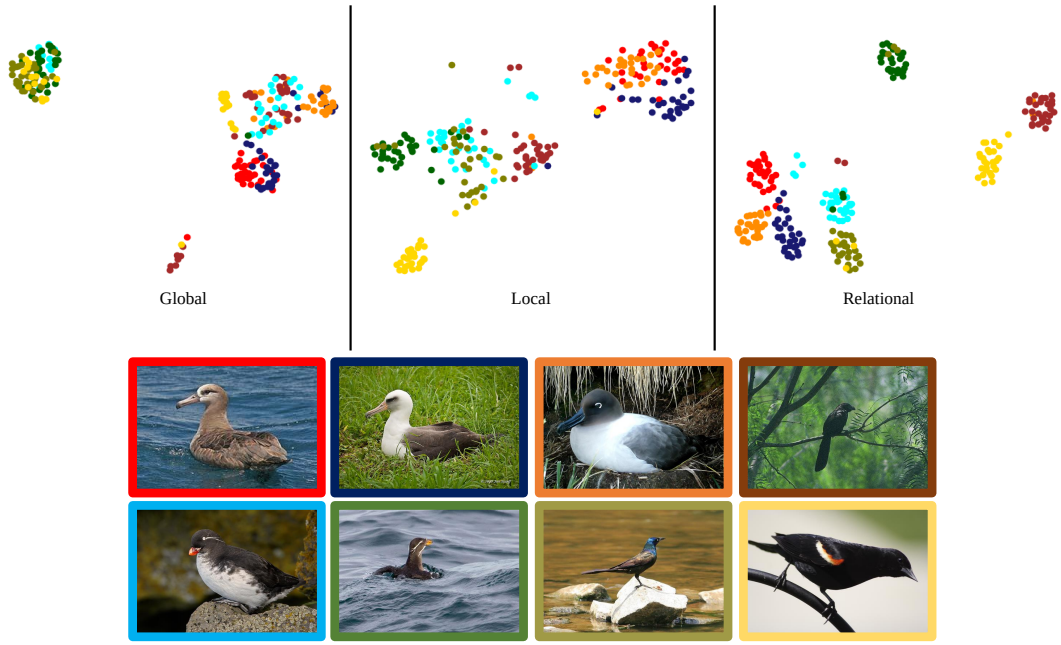

Figure 1: Top: Low dimensional embedding visualization of categories that are difficult to separate by global or local features alone, but can be separated using relational information. Bottom: Sample images from such categories. Colors indicate category memberships.

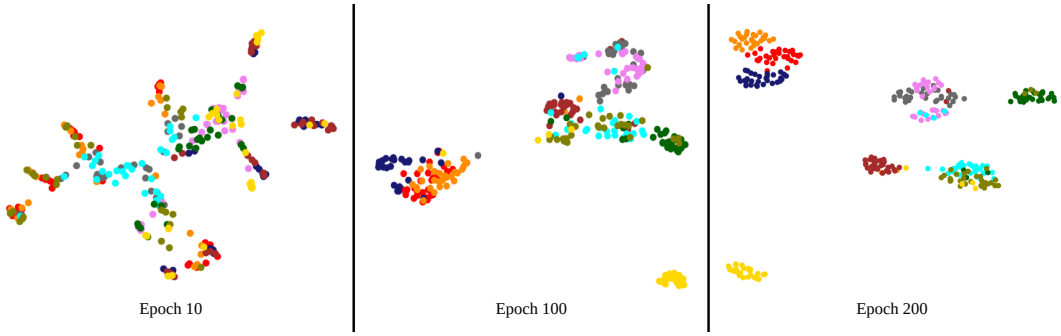

Figure 2: Low dimensional visualization of the relational representation ($\mathbf{r}$) space evolution across epochs. Colors indicate category memberships.

3 Additional notes on Relational Proxies

3.1 Pseudocode

Algorithm 1 provides the pseudocode for training our Relational Proxies model. We start by initializing a set of c learnable class-proxies $\{\mathbf{p}_1, \mathbf{p}_2, ... \mathbf{p}_c\}$. For each image $\mathbf{x}$, we obtain its global $\mathbf{z}_g$ and set $\mathbb{Z}_\mathbb{L}$ of local representations by propagating their corresponding views (obtained via cropping functions c_g and c_l) through a relation-agnostic encoder f (lines 10-11). We then realize the cross-view relational encoder ξ as a combination of the Attribute Summarization Transformer (AST) and the MLP ρ. The AST returns a summary of the local views $\mathbf{z}_\mathbb{L}$ (line 12). Using $\mathbf{z}_g$ and $\mathbf{z}_\mathbb{L}$, ρ computes the cross-view relation embedding $\mathbf{r}$ (line 13). Thereafter, all three representation of $\mathbf{x}$, *i.e.*, $\mathbf{z}_g$, $\mathbf{z}_l$ and $\mathbf{r}$ are used to condition the learning of the class proxies. The representations are incentivised to remain close to the proxy corresponding to their true class, while being distant from proxies of other classes (lines 15-19). How far the representation space deviates from this structural requirement is captured by computing the cross-entropic loss $\mathcal{L}_{\text{rproxy}}$. Minimizing $\mathcal{L}_{\text{rproxy}}$ thus has the effect of enforcing the representations to form a metric space (lines 23-27). Upon convergence, $\{\mathbf{p}_1, \mathbf{p}_2, ... \mathbf{p}_c\}$ serve as the set of Relational Proxies.

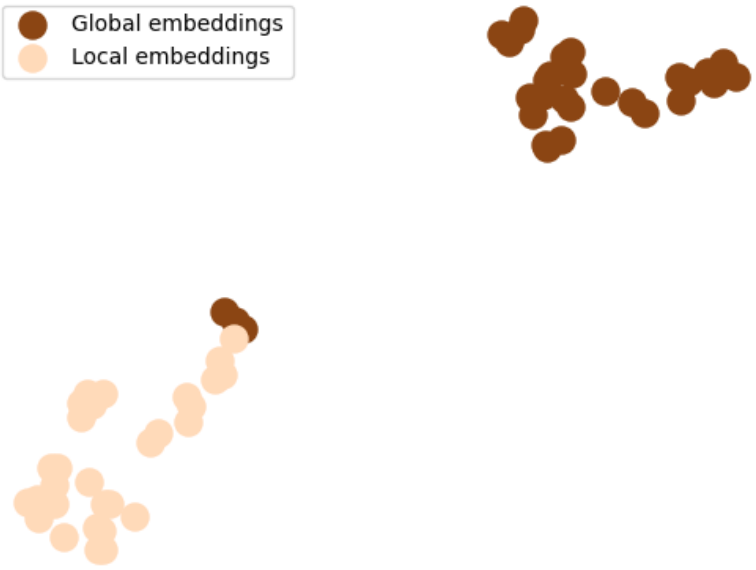

Figure 3: Low dimensional visualization of embeddings of global and local views for instances of a single class. The gap between the two clusters indicate the retention of relation-agnosticity even after the convergence of Relational Proxies, thereby supporting Lemma 3.

Figure 4: Male (left) and female (right) cardinals.

3.2 Cross-view relationships for intra-class variations

Figure 4 depicts the large variation in non-relational features like color and texture between male and female cardinals. Even though they belong to the same fine-grained category of cardinal birds, a model not accounting for the relationships between the individual local parts and the global view of the object, and hence not capturing the fine-grained geometric relationships, would not be able to map such significantly varying instances to the same neighborhood of the representation space. In such scenarios, the relational information becomes the only component that can be used to learn compact representations of categories with such large intra-class variations.

Algorithm 1: RELATIONAL-PROXIES: End-to-end training procedure for Relational Proxies.

Input : A set of images $\mathbb{X}$, their corresponding labels $\mathbb{Y}$, the number of fine-grained categories c, the number of epochs N, and the learning rate η.

Output : A relation agnostic-encoder f, a cross-view relation encoder ξ, and a set of c relational-proxies $\mathbb{P}$ corresponding to the unique labels in $\mathbb{Y}$.

`/* Initialize` c `learnable class-proxy vectors representing the labels in` $\mathbb{Y}$`. An image with label` $\mathbf{y}_i$ `has` $\mathbf{p}_i$ `as its corresponding class-proxy. */`

$\mathbb{P} \leftarrow \{\mathbf{p}_1, \mathbf{p}_2, ... \mathbf{p}_c\}$

for *epoch* $\leftarrow 1$ **to** N **do**

  $\mathcal{L}_{\text{rproxy}} \leftarrow 0$

  for $\mathbf{p} \in \mathbb{P}$ **do**

    $\psi^+ \leftarrow 0; \psi^- \leftarrow 0$

    for $\mathbf{x} \in \mathbb{X}$ **do**

      $\mathbf{g} \leftarrow c_g(\mathbf{x})$

      $\mathbb{L} \leftarrow \{\mathbf{l}_1, \mathbf{l}_2, ... \mathbf{l}_k\} \leftarrow c_l(\mathbf{x})$

      $\mathbf{z}_g \leftarrow f(\mathbf{g})$

      $\mathbb{Z}_{\mathbb{L}} \leftarrow \{\mathbf{z}_{l_1}, \mathbf{z}_{l_2}, ... \mathbf{z}_{l_k}\} \leftarrow \{f(\mathbf{l}) : \mathbf{l} \in \mathbb{L}\}$

      $\mathbf{z}_{\mathbb{L}} \leftarrow \text{AST}(\mathbb{Z}_{\mathbb{L}})$

      $\mathbf{r} \leftarrow \rho(\mathbf{z}_g, \mathbf{z}_{\mathbb{L}})$

      `// true proxy for` $\mathbf{x}$

      if $\mathbf{p} == \mathbf{p}^+$ **then**

        $\psi^+ \leftarrow \psi^+ + e^{-\alpha(s(\mathbf{z}_g, \mathbf{p}) - \delta)} + e^{-\alpha(s(\mathbf{z}_{\mathbb{L}}, \mathbf{p}) - \delta)} + e^{-\alpha(s(\mathbf{z}_g, \mathbf{p}) - \delta)}$

      `// negative proxies for` $\mathbf{x}$

      else

        $\psi^- \leftarrow \psi^- + e^{\alpha(s(\mathbf{z}_g, \mathbf{p}) + \delta)} + e^{\alpha(s(\mathbf{z}_{\mathbb{L}}, \mathbf{p}) + \delta)} + e^{\alpha(s(\mathbf{r}, \mathbf{p}) + \delta)}$

    $\psi^+ \leftarrow 1 + \psi^+$

    $\psi^- \leftarrow 1 + \psi^-$

    $\mathcal{L}_{\text{rproxy}} \leftarrow \mathcal{L}_{\text{rproxy}} - \frac{1}{c} \log\left(\frac{1}{\psi^+ \cdot \psi^-}\right)$

  $f \leftarrow f - \eta \nabla_f \mathcal{L}_{\text{rproxy}}$

  $\text{AST} \leftarrow \text{AST} - \eta \nabla_{\text{AST}} \mathcal{L}_{\text{rproxy}}$

  $\rho \leftarrow \rho - \eta \nabla_\rho \mathcal{L}_{\text{rproxy}}$

  for $\mathbf{p} \in \mathbb{P}$ **do**

    $\mathbf{p} \leftarrow \mathbf{p} - \eta \nabla_{\mathbf{p}} \mathcal{L}_{\text{rproxy}}$

References

[1] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeurIPS*, 2018.

[2] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. TransFG: A Transformer Architecture for Fine-grained Recognition. In *AAAI*, 2022.

[3] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform Manifold Approximation and Projection. *JOSS*, 2018.

[4] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing Properties of Vision Transformers. In *NeurIPS*, 2021.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[6] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization. In *MMM*, 2021.